Numerical techniques for the drift-diffusion semiconductor equations

Ferguson, R. C.

Award date:
1996

Awarding institution:
University of Bath

Numerical Techniques for the Drift-Diffusion Semiconductor Equations

submitted by

R.C. Ferguson

for the degree of PhD

of the

University of Bath

1996

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Summary

In this thesis we consider the numerical treatment of the steady state drift-diffusion semiconductor system. This system is composed of three coupled nonlinear elliptic equations which model the behaviour of the electrostatic potential and the electron and hole quasi-Fermi potentials of a semiconductor device. The analytic solutions can only be found in quite simple situations; instead we focus on how to go about finding accurate numerical solutions quickly and efficiently.

We apply the finite element method to the semiconductor system and show that Newton's method converges for sufficiently small voltage. Experiments show that the convergence ball of Newton's method is small, but can be extended by use of a continuation scheme. Instead we propose and analyse a decoupling method based on the mesh points of the discretisation rather than the partial differential equations. Experiments show that this alternative method does indeed extend the range of voltages we can find a solution for.

The solutions of the semiconductor system contain both interior layers and geometric boundary singularities which require appropriately graded meshes for accurate approximation. Since these irregularities are very complex and their precise position cannot be determined *a priori*, a mesh refinement strategy based on *a posteriori* error estimates is needed. In this thesis we derive *a posteriori* error estimates for a reduced class of problems and a theoretically justified efficient method of implementation which resolves the nonlinearity on a coarse mesh and then computes a sequence of corrections by solving linear problems on successively finer grids. We illustrate the use of these schemes on a number of typical semiconductor device models and demonstrate that we are able to capture important qualitative features of the solutions accurately and efficiently.

Acknowledgements

Thanks are due to my supervisor Ivan Graham, without his constant support, patience and unending enthusiasm this thesis would not have been possible. Special thanks should also go to John Martin, Adrian Hill and Mark Groves who have helped and inspired me in my efforts.

My colleagues in the School of Mathematical Science and the Student's Union have helped make my time in Bath particularly enjoyable. I am grateful to my office mates for being so indulgent and putting up with my whims. I should like to thank my family for their constant support.

I am grateful to EPSRC for providing financial support for this project.

> There is no abstract art.
> You must always start with something.
> Afterwards you can remove all traces of reality.
> *Pablo Picasso*

Contents

List of Figures

Chapter 1

Semiconductor Device Modelling, an Overview

1.1 Introduction

Semiconductors are widely used in today's technology. The rapid changes in electrical equipment means that semiconductors have had to develop and shrink in size. In a small device it is extremely difficult for the engineer to satisfactorily predict the consequences of slight changes in layout and design without accurate numerical simulation packages.

Numerical analysis can help with the three distinct phases of development which any new device goes through. Firstly the engineer needs to choose how to lay out the device and which of the many processes should be used to manufacture the semiconductor. In the second phase the proposed device needs to be optimised. here numerical simulations are needed to understand the special effects of the device and also to suggest experiments to test its behaviour. Finally numerical modelling is needed during the manufacture of the device, for instance it is necessary to know how the performance of the device is affected by problems with the manufacturing process. why these problems occur and how they can be rectified.

Semiconductor simulation and design needs accurate numerical tools specially designed for the semiconductor system. this is where we hope this thesis will make a contribution.

1.2 Review of Physics

This section contains a very brief look at semiconductors. It ignores many of the finer and more complicated details which are important if one wishes to understand semiconductors properly, these will be irrelevant in the context of this thesis. A full explanation of semiconductors can be found in books on solid state theory (e.g. Sze [67]), a particularly nice introduction to the subject can be found in Sparkes [65].

A semiconductor is a material which has an electrical conductivity significantly greater than an insulator, but smaller than a conductor. This can be seen by comparing the concentration of conducting electrons: a conductor (e.g. copper) has a concentration of the order 10^{22} cm^{-3}, an insulator (e.g. diamond) has a concentration of order 10^{3} cm^{-3}, but a semiconductor will typically have a concentration of order 10^{10} cm^{-3}. Common semiconductors are silicon and germanium. This thesis will concentrate on silicon (the preferred semiconductor for most uses), the discussion will apply to other semiconductors, but with different constants.

Silicon atoms form a regular tetrahedral structure, each atom has four nearest neighbours to which it is bound by covalent bonds. The four electrons in the outer shell of each silicon atom are shared with its neighbours, so each bond can be thought of as containing two electrons. At low temperatures all these electrons are held firmly in place and the material acts like an insulator.

If the temperature of the silicon structure is raised then thermal energy is introduced into the silicon. This energy will not be distributed evenly. Those electrons which receive enough energy can break free of their bonds and become conducting **electrons**. As the temperature is raised the silicon acts more like an electrical conductor.

The free electrons can move within the spaces between the bonds and are said to conduct negative charge. The gap left in the bond is called a **hole**, these holes can move within the bonds and, in effect, carry a positive charge (with equal magnitude to the negative charge of an electron). These holes and electrons can only move about for a limited amount of time, eventually the holes and electrons recombine, but not necessarily with the electron or hole that first formed the original electron-hole pair. This process is called electron/hole generation/recombination.

The effect of this process is to produce sufficient particles of each type, both of which are capable of conducting electricity. In a pure semiconductor, even at quite

high temperatures, there are relatively few conducting holes and electrons. **Doping** introduces more and allows the engineer to control the way the semiconductor conducts electricity.

If some of the silicon atoms are replaced by atoms with five electrons in their outer shell (for example arsenic or phosphorus), then there will be extra non-bonded electrons in the semiconductor (one for each atom introduced). Such atoms are called **donors**. A semiconductor doped with donor atoms does not require high temperatures to conduct electricity.

It is also possible to add atoms with only three electrons in their outer shell (e.g. aluminium or boron), these atoms are called **acceptors**. The acceptor atoms have a shortage of electrons for bonding and thus introduce holes into the structure.

A semiconductor doped with donors is called an **n-type** semiconductor, while one with acceptor atoms is called a **p-type** semiconductor. The way in which a semiconductor will conduct electricity can be altered by combining n- and p-type semiconductors.

As an example consider the **PN diode**. A PN diode is a single crystal of semiconductor which has a region of p-type semiconductor next to a region of n-type semiconductor. The transition from p-type to n-type material is called a **PN junction**. PN diodes have the property that large currents can only pass in one direction through the device. If the n-type side of the PN junction has a more negative voltage applied to it than the p-type region then the semiconductor is said to be in **forward bias** and a large current can flow. If the n-type region is more positive then the device is in **reverse bias** and the semiconductor will not conduct well, a small current may flow if there is a large enough difference between the applied voltages. A PN diode and its properties is shown in Figure 1-1.

1.3 The Drift-Diffusion Equations

The basic mathematical model which is commonly used for analysis and simulation of a semiconductor device is the set of equations:

$$\Delta\psi = \frac{q}{\epsilon}(n - p - d), \tag{1.3.1}$$

$$\nabla . J_n - q\frac{\partial n}{\partial t} = qr. \tag{1.3.2}$$

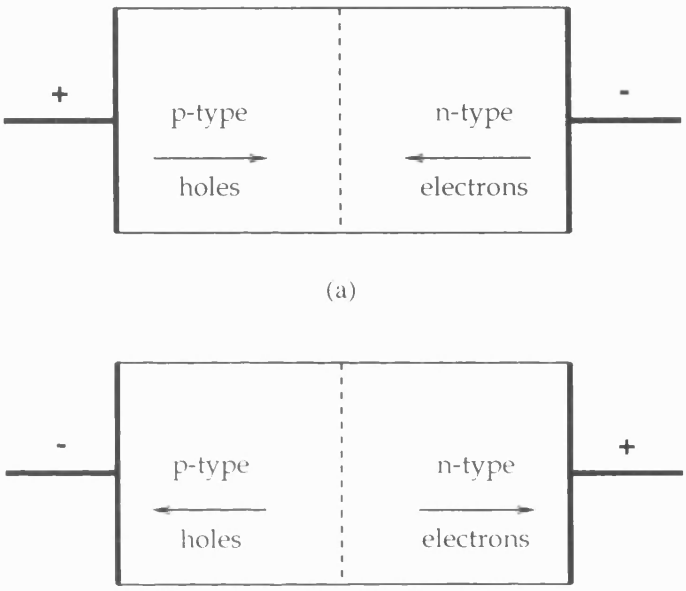


Figure 1-1: A PN diode under various applied voltages. (a) shows a PN diode with positive voltage applied to the contact at the p-type region and negative voltage applied at the n-type region, this is the forward bias case. Large current flows. (b) shows the PN diode with the voltages reversed, the device is in reverse bias and only a small current can flow.

$$\nabla . J_p + q\frac{\partial p}{\partial t} = -qr, \tag{1.3.3}$$

$$J_n = q\mu_n(U_T\nabla n - n\nabla\psi), \tag{1.3.4}$$

$$J_p = -q\mu_p(U_T\nabla p + p\nabla\psi). \tag{1.3.5}$$

(1.3.1)-(1.3.5) are derived from Maxwell's equations under certain assumptions, the details of which can be found in many semiconductor texts, see for example [22], [50] and [62]. (1.3.1)-(1.3.5) are almost exactly the set of partial differential equations (PDEs) first used to model semiconductors in 1950 by Van Roosbroeck, [68]. Details of other semiconductor models can be found in [62].

In (1.3.1)-(1.3.5) the dependent variables to be found are the **electrostatic potential** ψ, the **electron concentration** n and the **hole concentration** p. It is important in many applications to calculate the electron and hole current densities J_n and J_p, but this will not be one of the aims of this thesis. (1.3.1)-(1.3.5) ignore the important affect of temperature on the semiconductor device. Here temperature will be assumed to be an externally defined (positive) constant, in some other models temperature is assumed to be variable and is governed by an additional equation.

It is assumed that the voltages applied to the contacts of the device, ψ, n, p, J_n and J_p are all time-independent. As a consequence set

$$\frac{\partial n}{\partial t} = \frac{\partial p}{\partial t} = 0$$

in (1.3.2) and (1.3.3). The steady state solutions computed can be thought of as modelling the long term performance of the device, see [50].

In (1.3.1)-(1.3.5) μ_n and μ_p are the electron and hole carrier mobilities, q is the elementary charge, ϵ is the absolute permittivity of the semiconductor material and U_T is the thermal voltage. Typical values of these quantities, taken from [50], are given in Table 1.1.

r in (1.3.2) and (1.3.3) is the generation/recombination term, this is a function of n and p and describes the balance of generation and recombination of electrons and holes. The generation of hole and electron pairs occurs when $r > 0$ and recombination occurs when $r < 0$. There are many different models for r, see for example [50], but for the purposes of this thesis we restrict our attention to the commonly used Shockley-Read-

Hall recombination rate:

$$\frac{np - n_i^2}{\tau_p(n + n_i) + \tau_n(p + n_i)}. \tag{1.3.6}$$

τ_n and τ_p in (1.3.6) are the electron and hole carrier lifetimes and n_i is the intrinsic concentration, values of which are given in Table 1.1. To simplify the model set $\tau_n = \tau_p = \tau = 1 \times 10^{-6}$, this is a common assumption for many semiconductor simulations, although a more realistic value for τ_p is 1×10^{-5}.

Quantity	Symbol	Typical value
Electron carrier mobility	μ_n	1000 $\text{cm}^2\text{V}^{-1}\text{s}^{-1}$
Hole carrier mobility	μ_p	1000 $\text{cm}^2\text{V}^{-1}\text{s}^{-1}$
Elementary charge	q	$1.602 \times 10^{-19}\text{As}$
Absolute permittivity in a vacuum	ϵ_ν	$8.854 \times 10^{-14}\text{AsV}^{-1}\text{cm}^{-1}$
Absolute permittivity of silicon	ϵ_s	11.7 ϵ_ν
Absolute permittivity of silicon-dioxide	ϵ_{ox}	3.9 ϵ_ν
Thermal voltage	U_T	0.0258520 V
Electron carrier lifetime	τ_n	1×10^{-6}s
Hole carrier lifetime	τ_p	1×10^{-6}s
Intrinsic concentration	n_i	$1 \times 10^{17}\text{cm}^{-3}$
Maximum of doping profile	$\tilde{d}$	1×10^{17}
Device diameter	l	1×10^{-6}cm

Table 1.1: Typical values of constants appearing in the basic semiconductor model.

In equation (1.3.1) d is the doping profile of the device and depends on the type of semiconductor being studied. The doping profile of a device reflects the implantation of doping atoms into the semiconductor device (donor and acceptor atoms). The doping profile measures the concentration of these active doping atoms at each part of the device:

$$d = N_D - N_A,$$

where N_D denotes the concentration of electrically active donor atoms and N_A is the concentration of electrically active acceptor atoms.

A section of the device is called an n-type domain if the concentration of donors exceeds the concentration of acceptor atoms. $d > 0$ in an n-type domain. A p-type domain has $d < 0$ and is a region where the concentration of acceptors exceeds the concentration of donor atoms.

To simplify the modelling of a semiconductor device the doping profile is often assumed to be piecewise constant. Such an assumption is reasonable as the doping

profile of real devices varies slowly within a n- or p-type domain and fast at the junctions between the n- and p-type domains. Further details on doping profiles for semiconductor devices can be found in [50, Section 2.2].

The equations governing the steady state semiconductor model are very badly scaled. In addition n and p are typically of the order 10^{16} m^{-3}. This makes the equations very difficult to solve numerically. Some of these problems are relieved by scaling the spatial variable $\boldsymbol{x}$ by the characteristic device diameter l:

$$\tilde{\psi}(\boldsymbol{x}) = \psi(l\boldsymbol{x}), \quad \tilde{n}(\boldsymbol{x}) = n(l\boldsymbol{x}), \quad \tilde{p}(\boldsymbol{x}) = p(l\boldsymbol{x})$$

and then defining new variables ψ, n, p and d by:

$$\psi = \tilde{\psi}(\boldsymbol{x})/U_T, \; n = \tilde{n}/\tilde{d}, \; p = \tilde{p}/\tilde{d}, \; d = d/\tilde{d},$$

where $\tilde{d} = \max\{|d(\boldsymbol{x})| : \boldsymbol{x} \in \Omega\}$. With these new variables the scaled (and simplified) equations are:

$$-\lambda^2 \Delta\psi = p - n + d, \tag{1.3.7}$$

$$\mu_n \nabla.(\nabla n - n\nabla\psi) = \frac{l^2 r}{\tilde{d} U_T}, \tag{1.3.8}$$

$$\mu_p \nabla.(\nabla p + p\nabla\psi) = \frac{l^2 r}{\tilde{d} U_T}. \tag{1.3.9}$$

In (1.3.7), $\lambda = l^{-1}\sqrt{\epsilon U_T/q\tilde{d}}$ is called the **Debye length**.

To restrict the range of n and p it is often helpful to change from the charge concentrations n and p to the **quasi-Fermi** levels v and w. [This is only one of many possible changes of variables]. v is called the **electron quasi-Fermi** level and w the **hole quasi-Fermi** level. The change of variables is achieved by the transformation:

$$n = \frac{n_i}{\tilde{d}} \exp(\psi - v), \tag{1.3.10}$$

$$p = \frac{n_i}{\tilde{d}} \exp(w - \psi). \tag{1.3.11}$$

In order for this transformation to be valid it is assumed that the Boltzmann statistics hold for the carrier concentrations, see [62]. With the quasi-Fermi variables (1.3.7)-

(1.3.9) becomes:

$$-\lambda^2\Delta\psi + \delta^2\{\exp(\psi - v) - \exp(w - \psi)\} = d, \tag{1.3.12}$$

$$-\nabla.(\exp(\psi - v)\nabla v) = \sigma\rho_v r, \tag{1.3.13}$$

$$\nabla.(\exp(w - \psi)\nabla w) = \sigma\rho_w r. \tag{1.3.14}$$

In the above $\rho_v = 1/\mu_n$, $\rho_w = 1/\mu_p$, $\delta^2 = n_i/\tilde{d}$, $\sigma = l^2/n_i U_T$ and r, the generation/recombination rate, is given by:

$$r = \frac{n_i}{\tau}\frac{\exp(w - v) - 1}{(\exp(\psi - v) + \exp(w - \psi) + 2)}. \tag{1.3.15}$$

Equations (1.3.12)-(1.3.15) are known as the **drift-diffusion semiconductor equations**, but as they are the only set of equations used to model the full semiconductor system in this thesis they will often be referred to as **the** semiconductor equations. (1.3.12) is known as the **Poisson Boltzmann equation** or just the **potential equation**.

To complete the study of the equations modelling a semiconductor device it is necessary to specify the boundary conditions associated with ψ, v and w. The boundary of a device can usually be split into two parts: a part corresponding to real physical boundaries and a part corresponding to artificial boundaries introduced to separate adjacent devices or to cut off a device to simplify the simulation.

The artificial boundary is important for devices embedded in integrated circuits, the MOSFET discussed in Section 6.3 is one such device. On artificial boundaries it is usual to assume homogeneous Neumann boundary conditions for ψ, v and w and also to impose some interface conditions.

The physical boundary corresponds to semiconductor-oxide interfaces, insulated segments and metal contacts. There are two different types of metal contacts used in modern devices: Ohmic and Schottky contacts. For simplicity we restrict our attention to Ohmic contacts. For an applied voltage of V_a at the contact, O, the boundary condition for the quasi-Fermi variables are:

$$v|_O = w|_O = V_a/U_T, \tag{1.3.16}$$

$$\left[\delta^2\{\exp(\psi - v) - \exp(w - \psi)\} - d\right]|_O = 0. \tag{1.3.17}$$

(1.3.17) is the **vanishing space charge** condition at the contact and corresponds to the requirement that no new charge is introduced into the device. (1.3.17) is equivalent to the boundary condition:

$$\psi|_O = V_{bi} + V_a/U_T \tag{1.3.18}$$

where the **built in voltage**, V_{bi}, is given by:

$$V_{bi} = \sinh^{-1}\left(\frac{d|_0}{2\delta^2}\right). \tag{1.3.19}$$

At the insulating segments of the device we assume homogeneous Neumann boundary conditions. Things are more complicated at a semiconductor-oxide interface and, since such interfaces only occur in Metal Oxide Semiconductors (MOS), we deal with these boundaries when they arise in Section 6.3.

In summary, for most of the semiconductor devices we deal with the boundary of the device can be split into two disjoint sets - a union of Dirichlet boundaries and a union of Neumann boundaries. The Dirichlet boundaries correspond to the contacts of the device and have boundary conditions given by (1.3.16) and (1.3.18). On the Neumann boundaries (corresponding to insulating segments or artificial boundaries) the boundary conditions are taken as homogeneous Neumann. Further details of the boundary conditions can be found in [50] and [62].

1.4 Review of Numerical Methods applied to the Semiconductor Equations

In this section we take a brief look at previous work on the numerical treatment of the drift-diffusion semiconductor equations. The range of methods used makes it impossible to give a complete literature survey, instead we give an overview of recent trends.

The first step in the numerical solution of the drift-diffusion semiconductor equations is the choice of discretisation method. Here we work with the standard finite element method, but the finite difference and finite volume (or box) methods are also popular. Of particular importance is a hybrid finite element scheme first introduced by Brezzi *et al.* [16], [17]. This hybrid scheme exhibits current conservation properties, which is of interest to the engineer in practical applications. The scheme acts on the exponential terms appearing in the current continuity equations (1.3.13), (1.3.14) and,

in one dimension, is equivalent to the finite element method we consider. Although we do not consider such a scheme in this thesis the methods described herein can be extended to include it.

Of the many numerical methods used to solve the discretised semiconductor equations the most popular are Gummel's method and Newton's method. Gummel's method is the name given to a general family of nonlinear Gauss-Seidel methods and one particular variant will be studied in Chapter 3. Newton's method applied to the semiconductor system will be considered in Chapter 2.

Much of the current work on semiconductor modelling is concerned with the efficient solution of the discretised equations. For example issues associated with the solution of the equations on parallel machines are considered in [23] and [55]. Multigrid methods have also been applied to the semiconductor system in, for example [26], [52] and [59]. In addition there is a considerable amount of recent work focused on adaptive methods. These adaptive methods generally fall into four categories:

1. Grid generation and adaption based on the doping profile (e.g. [24]).

2. Refinement based on the change in the gradient of the finite element solutions [known as gradient smoothing] (e.g. [48]).

3. Adaptive procedures based on the potential equation (e.g. [47]).

4. Refinement based on all three semiconductor equations (e.g. [18]).

Most of this work is non-rigorous or is based on *a priori* knowledge, in Chapter 5 we consider a rigorous adaptive procedure for solving the semiconductor equations based on the potential equation. We give various experiments to show that the method is capable of accurately capturing the features of the asymptotic solutions.

1.5 What This Thesis Contains

As we have seen the drift-diffusion equations are three coupled nonlinear elliptic equations. The analytic solutions can only be found in quite simple situations, and, instead we focus on how one could go about finding accurate numerical solutions quickly and efficiently. In this thesis the emphasis will be placed on finding new or improved numerical routines that extend the convergence ball of the method or significantly reduce the

amount of effort needed to find accurate solutions. We will not be so concerned with the optimality of the code used to implement the methods and much of the implementation details will not be covered in this thesis.

In Chapter 2 we study the convergence of Newton's method applied to the discretised semiconductor system. Results show that Newton's method, where the inner linear systems are solved using a block Gauss-Seidel iteration, only converges for small bias when applied to the semiconductor system. We give details of a method combining Newton's method with a continuation scheme which significantly extends the convergence ball.

In Section 3.3.3 of Chapter 3 we study "Gummel's Method", arguably the most popular choice of method for solving the semiconductor system. Results are given which show that the method converges for small reverse bias and much larger forward bias.

Our work on the Gummel and Newton methods suggest that the failure of these methods to converge for reasonable applied voltage might be due to the way we approach the solution of the equations. It has been suggested that the coupling between the equations is stronger than the coupling between the values of the solution to a single PDE at the mesh points. The Gummel and Newton type methods used in device modelling do not take full account of the coupling between the equations. For instance the Jacobia arising in Newton's method have a 3×3 block structure and it is usual to use a block Jacobi or block Gauss-Seidel iteration to solve these linear systems. This leads to successive solutions of each PDE in turn and neglects the coupling between the PDEs at each step. We propose an alternative method originally introduced by Bank *et al.* in [8], but only ever studied empirically, which aims to preserve the coupling between the equations. This method solves for all the unknowns at each of the mesh points of the discretisation in turn, i.e. we use a Jacobi iteration to solve for ψ, v and w at the first mesh point, then the second, etc. and repeat until convergence. It is shown in Chapter 3 that this method applied to the semiconductor system does indeed converge. Numerical simulations also show that the applied voltages we can solve for is significantly extended when compared to Newton's or Gummel's method. To the best of our knowledge this is the first rigorous result concerning this alternative method.

The rest of the thesis is concerned with adaptive and multilevel methods. In Chapters 4 and 5 we consider the efficient accurate solution of a single semilinear PDE. We study semilinear equations as the potential equation in the semiconductor system is of

this form and also because Gummel's method leads naturally to a system of semilinear equations. Typical adaptive methods for nonlinear problems solve, to full accuracy, a nonlinear problem for each triangulation before computing an error estimate and refining the grid. Since a typical refinement process can involve refining a number of triangulations this results in a lot of wasted effort. In Chapter 4 we propose a method that considerably reduces this effort by solving a nonlinear problem on the coarsest mesh and then one linear problem for each of the finer meshes. We call the method the defect correction method. Under the assumption that a suitable *a priori* determined mesh sequence is used, it is shown that this method is well defined and has an error estimate that is essentially the same as that satisfied by the standard finite element solution. In fact, neglecting higher order terms, the error in the defect correction solution is asymptotically bounded from above and below by the error in the standard finite element solution on the same mesh.

The solutions to the semiconductor system contain both interior layers and geometric boundary singularities which require appropriately graded meshes for their accurate approximation. These singularities are very complex and the precise position of the interior layers is quite delicate, it is not possible to derive suitable meshes *a priori* and a mesh refinement process based on *a posteriori* error estimation is necessary. In Chapter 5 we give *a posteriori* error estimates for general semilinear equations on polygonal domains which works well under extreme parameter ranges and in the presence of geometric singularities. By considering model semilinear semiconductor problems with known asymptotic solutions we demonstrate that an adaptive procedure based on the *a posteriori* error estimates is capable of finding accurate finite element solutions displaying the correct asymptotic features. The constants appearing in the *a posteriori* error estimates are estimated and compared to the theoretical bounds. It is shown heuristically that the true values of the constants are likely to be closer to the estimated constants, rather than to the theoretical bounds. We also test the efficiency of our adaptive procedure and show that the method is close to optimal.

In Chapter 6 we combine the *a posteriori* error estimate with the defect correction method to find accurate finite element solutions cheaply. The work here differs from that in Chapter 4 as we use adaptively determined meshes, rather than the *a priori* meshes of the theory. We demonstrate that the finite element solutions produced for a model problem have the correct asymptotic features and show that the solutions are cheap

to calculate. Finally we apply the adaptive procedure to a simplified MOSFET diode with non-zero applied voltage and demonstrate that our *a posteriori* error estimates are capable of capturing the features of a problem which can not naturally be written in semilinear form.

Chapter 2

Convergence of Newton's Method for the 2D Semiconductor System

Newton's method and its variants are widely used in semiconductor device modelling, see for example [8], [39] and [33]. In this chapter we aim to show that Newton's method applied to the finite element system arising from the discretisation of the drift-diffusion semiconductor equations does indeed converge for sufficiently small voltage.

At the end of the chapter numerical results showing the performance of Newton's method applied to the finite element discretisation of the one dimensional semiconductor system are given. The results show that the method converges for a similar range of applied voltages to that of Gummel's method (discussed in Chapter 3). To extend the convergence ball of the method we also give results for a procedure combining Newton's method with a continuation scheme. It is shown that the method significantly increases the range of applied voltages it is possible to solve for.

Results obtained in this chapter will be extensively used in later chapters, in particular we will show that for sufficiently small applied voltage

- The finite element solution exists.
- The Jacobian matrix of the finite element discretisation of semiconductor system is non-singular at the true finite element solution.
- The Jacobian matrix of the finite element system is continuous with respect to both the finite element solution and the applied voltage.

The proof of convergence will require matrix theory and for this we refer the reader to Appendix A.

2.1 The General Semiconductor System

Here we are considering the full drift-diffusion semiconductor system given below. We aim to prove that Newton's method applied to the finite element discretisation of this system converges.

We approximate the solutions ψ, v and w of the system:

$$-\lambda^2\Delta\psi + \delta^2\{\exp(\psi - v) - \exp(w - \psi)\} - d = 0, \tag{2.1.1}$$

$$-\nabla.(\exp(\psi - v)\nabla v) - \sigma\rho_v r(\psi, v, w) = 0, \tag{2.1.2}$$

$$-\nabla.(\exp(w - \psi)\nabla w) + \sigma\rho_w r(\psi, v, w) = 0. \tag{2.1.3}$$

In the above λ, δ, σ, ρ_v and ρ_w are considered to be positive constants and d is the piecewise continuous doping profile. ψ is called the electrostatic potential of the device, v and w are "quasi-Fermi potentials" which determine the electron and hole densities. In our model we use the Shockley-Read-Hall generation/recombination rate, r, given by:

$$r(\psi, v, w) = \frac{\exp(w - v) - 1}{\exp(w - \psi) + \exp(\psi - v) + 2} \tag{2.1.4}$$

We consider solving (2.1.1)-(2.1.3) in a connected polygonal domain $\Omega \subset \mathbb{R}^2$ with boundary $\partial\Omega$. We apply the following boundary conditions to the system:

Split $\partial\Omega$ into two parts: a Dirichlet part, $\partial\Omega_D$, corresponding to the contacts of the semiconductor device and a Neumann part, $\partial\Omega_N$. We assume that $\partial\Omega_D$ and $\partial\Omega_N$ are composed of straight line segments of $\partial\Omega$ and further that

$$\partial\Omega_D \cap \partial\Omega_N = \emptyset. \qquad \partial\Omega_D \cup \partial\Omega_N = \partial\Omega. \tag{2.1.5}$$

On $\partial\Omega_N$ we require

$$\frac{\partial\psi}{\partial n} = \frac{\partial v}{\partial n} = \frac{\partial w}{\partial n} = 0. \tag{2.1.6}$$

Define on $\partial\Omega_D$ the built in voltage of the device β:

$$\beta = \sinh^{-1}\left(\frac{d|_{\partial\Omega_D}}{2\delta^2}\right). \tag{2.1.7}$$

We note that β is piecewise continuous on $\partial\Omega_D$. For a given α, piecewise continuous on $\partial\Omega_D$, we require ψ, v and w to satisfy the Dirichlet boundary conditions:

$$v|_{\partial\Omega_D} = w|_{\partial\Omega_D} = \alpha \tag{2.1.8}$$

and

$$\psi|_{\partial\Omega_D} = \alpha + \beta. \tag{2.1.9}$$

Here α is the function comprised of the scaled voltages applied at the contacts which make up $\partial\Omega_D$. Zero α will correspond to zero applied voltage.

2.2 The Finite Element System

To approximate the weak solutions of (2.1.1)-(2.1.3) with the given boundary conditions we use the piecewise linear finite element method. Define $\mathcal{T}_h = \{\mathrm{T}\}$ to be a triangulation of Ω. We make the following assumptions on the triangulation:

(M1) $\Omega = \bigcup_{\mathrm{T}_k \in \mathcal{T}_h} \mathrm{T}_k$.

(M2) If $\mathrm{T}_1, \mathrm{T}_2 \in \mathcal{T}_h$, $\mathrm{T}_1 \neq \mathrm{T}_2$, then T_1 and T_2 are either disjoint or have a vertex in common, or an edge in common.

(M3) For each interior triangle edge, the sum of the two angles opposite it should be no greater than π. For a triangle edge on a Neumann boundary the angle opposite should be no greater than $\pi/2$.

(M4) No edge of a triangle on the boundary of Ω has both Dirichlet and Neumann boundary conditions defined on it.

Remark 2.2.1 *The mesh condition* **(M3)** *arises in [45] and will be needed if it is required that the matrices in the finite element discretisation have positive inverse.*

Each vertex of a triangle in our triangulation will be called a mesh point and will typically be denoted by p or q. $\mathcal{N}$ will be the set of mesh points not on the Dirichlet

boundary, $\partial\Omega_D$, and $\mathcal{N}_D$ is the set of mesh points on $\partial\Omega_D$. We let $[\mathcal{N}]$ denote the space of all real vectors which have a unique entry for each mesh point in the set $\mathcal{N}$, $[\mathcal{N}]^3$ is the space of all real vectors in $[\mathcal{N}] \times [\mathcal{N}] \times [\mathcal{N}]$.

We introduce the space of piecewise linear functions, $\mathcal{V}_h$, based on our triangulation $\mathcal{T}_h$ of Ω. Define the hat functions, ϕ_p, $p \in \mathcal{N} \cup \mathcal{N}_D$ to satisfy $\phi_p(q) = \boldsymbol{\delta}_{pq}$, where $\boldsymbol{\delta}_{pq}$ is the Kronecker delta. A basis for $\mathcal{V}_h$ is $\{\phi_p : p \in \mathcal{N} \cup \mathcal{N}_D\}$.

Unless we indicate otherwise we shall use the uniform norm on $\mathbb{R}^n$ and $\mathcal{B}(\boldsymbol{X}, \alpha)$ will be the open ball centred at $\boldsymbol{X}$ with radius α.

The standard finite element method for (2.1.1)-(2.1.3) is to seek Ψ, V and W in $\mathcal{V}_h$, satisfying the boundary conditions (2.1.8) and (2.1.9), such that

$$\lambda^2(\nabla\Psi, \nabla\phi_p) + (\delta^2\{\exp(\Psi - V) - \exp(W - \Psi)\} - d, \phi_p) = 0, \tag{2.2.10}$$

$$(\exp(\Psi - V)\nabla V, \nabla\phi_p) - (\sigma\rho_v r(\Psi, V, W), \phi_p) = 0, \tag{2.2.11}$$

$$(\exp(W - \Psi)\nabla W, \nabla\phi_p) + (\sigma\rho_w r(\Psi, V, W), \phi_p) = 0, \tag{2.2.12}$$

where p ranges over the set $\mathcal{N}$.

In fact we will consider a slightly modified scheme to (2.2.10)-(2.2.12) where we replace the second terms of (2.2.10), (2.2.11) and (2.2.12) by their mass lumped versions. As discussed in Appendix B this is obtained by replacing a term of the form (f, g) by its discrete counterpart $\langle f, g\rangle$, where

$$\begin{aligned} \langle f, g\rangle &:= \sum_{\mathrm{T}\in\mathcal{T}_h} \frac{1}{3}\mathcal{A}(\mathrm{T}) \sum_{p\in\mathrm{T}} (fg)(p) \\ &=: \sum_{p\in[\mathcal{N}]\cup[\mathcal{N}_D]} w_p\ (fg)(p). \end{aligned} \tag{2.2.13}$$

Here $\mathcal{A}(\mathrm{T})$ denotes the area of triangle $\mathrm{T} \in \mathcal{T}_h$ and w_p is a third the sum of the areas of all triangles meeting at the mesh point p.

Thus the mass lumped system is :

$$\lambda^2(\nabla\Psi, \nabla\phi_p) + \langle\delta^2\{\exp(\Psi - V) - \exp(W - \Psi)\} - d, \phi_p\rangle = 0. \tag{2.2.14}$$

$$(\exp(\Psi - V)\nabla V, \nabla\phi_p) - \langle\sigma\rho_v r(\Psi, V, W), \phi_p\rangle = 0. \tag{2.2.15}$$

$$(\exp(W-\Psi)\nabla W, \nabla\phi_p) \; + \; \langle \sigma\rho_w r(\Psi, V, W), \phi_p\rangle = 0, \tag{2.2.16}$$

where p ranges over $\mathcal{N}$.

Since Ψ, V and W are members of $\mathcal{V}_h$ and satisfy the boundary conditions (2.1.6), (2.1.8) and (2.1.9), we may write:

$$\Psi = \sum_{p\in\mathcal{N}} \Psi_p\phi_p + \sum_{q\in\mathcal{N}_D} (\alpha_q + \beta_q)\phi_q, \tag{2.2.17}$$

$$V = \sum_{p\in\mathcal{N}} V_p\phi_p + \sum_{q\in\mathcal{N}_D} \alpha_q\phi_q, \tag{2.2.18}$$

$$W = \sum_{p\in\mathcal{N}} W_p\phi_p + \sum_{q\in\mathcal{N}_D} \alpha_q\phi_q. \tag{2.2.19}$$

Where $\alpha_q = \alpha(q)$, $\beta_q = \beta(q)$, for q in $\mathcal{N}_D$ and α and β are as given in (2.1.7) and (2.1.8).

Thus the computation of Ψ, V, W is equivalent to the problem of finding the vector of unknowns $\boldsymbol{X}:= \left(\boldsymbol{\Psi}^T, \boldsymbol{V}^T, \boldsymbol{W}^T\right)^T \in [\mathcal{N}]^3$ which appears in (2.2.17)-(2.2.19), for a given $\boldsymbol{\alpha} \in [\mathcal{N}_D]$ (the other parameter, $\boldsymbol{\beta}$, is assumed to be given *a priori*). We shall consider the behaviour of solutions of this system with respect to variations in $\boldsymbol{\alpha}$, providing a discrete version of well-known results in the continuous case (e.g. [51]).

Define $\tilde{\boldsymbol{\Psi}} \in [\mathcal{N}] \times [\mathcal{N}_D]$ to be the extended vector, including the values of Ψ at the mesh points on the Dirichlet boundary. $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{W}}$ are defined analogously. The problem of finding $\boldsymbol{X}$ may be written in the more compact form:

$$\mathbf{F}(\boldsymbol{X}, \boldsymbol{\alpha}) = \mathbf{0} \tag{2.2.20}$$

where the function $\mathbf{F}:= \left(\mathbf{F}_1^T, \mathbf{F}_2^T, \mathbf{F}_3^T\right)^T : [\mathcal{N}]^3 \times [\mathcal{N}_D] \to [\mathcal{N}]^3$ is defined as follows:

$$\mathbf{F}_1(\boldsymbol{X}, \boldsymbol{\alpha}) = \lambda^2 \tilde{A}(\mathbf{0})\tilde{\boldsymbol{\Psi}} + \boldsymbol{e}(\boldsymbol{\Psi} - \boldsymbol{V}) - \boldsymbol{e}(\boldsymbol{W} - \boldsymbol{\Psi}) - \boldsymbol{d}, \tag{2.2.21}$$

$$\mathbf{F}_2(\boldsymbol{X}, \boldsymbol{\alpha}) = \tilde{A}\left(\tilde{\boldsymbol{\Psi}} - \tilde{\boldsymbol{V}}\right)\boldsymbol{V} - \rho_v \boldsymbol{r}(\boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W}). \tag{2.2.22}$$

$$\mathbf{F}_3(\boldsymbol{X}, \boldsymbol{\alpha}) = \tilde{A}\left(\tilde{\boldsymbol{W}} - \tilde{\boldsymbol{\Psi}}\right)\boldsymbol{W} + \rho_w \boldsymbol{r}(\boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W}). \tag{2.2.23}$$

The matrices and vectors are defined by:

$$\begin{aligned}
\tilde{A}(\boldsymbol{B})_{pq} &= \left(\exp\left(\sum_{r\in\mathcal{N}\cup\mathcal{N}_D} B_r\phi_r\right)\nabla\phi_p,\nabla\phi_q\right), \quad p\in\mathcal{N},\ \ q\in\mathcal{N}\cup\mathcal{N}_D, \\
&\qquad\qquad \boldsymbol{B}\in[\mathcal{N}]\times[\mathcal{N}_D], \\
e(\boldsymbol{B})_p &= w_p\delta^2\exp(B_p), \quad p\in\mathcal{N},\ \boldsymbol{B}\in[\mathcal{N}], \\
d_p &= w_p d(p), \quad p\in\mathcal{N}, \\
r(\boldsymbol{\Psi},\boldsymbol{V},\boldsymbol{W})_p &= w_p\sigma r(\Psi_p,V_p,W_p), \quad p\in\mathcal{N}.
\end{aligned}$$

In the above, w_p is one third of the areas of all triangles meeting at mesh point p, as defined implicitly in (2.2.13).

2.3 Newton's Method

In this section we shall prove that for fixed $\boldsymbol{\alpha}$, $\|\boldsymbol{\alpha}\|$ sufficiently small, Newton's method converges when applied to the system (2.2.20). For the proof we need to consider the Jacobian, J, of (2.2.20) with respect to $\left(\boldsymbol{\Psi}^T,\boldsymbol{V}^T,\boldsymbol{W}^T\right)^T$. An elementary but tedious calculation shows that (in block notation):

$$\mathrm{J}\left(\left(\boldsymbol{\Psi}^T,\boldsymbol{V}^T,\boldsymbol{W}^T\right)^T,\boldsymbol{\alpha}\right) = \begin{bmatrix} \mathrm{J}_{11} & \mathrm{J}_{12} & \mathrm{J}_{13} \\ \mathrm{J}_{21} & \mathrm{J}_{22} & \mathrm{J}_{23} \\ \mathrm{J}_{31} & \mathrm{J}_{32} & \mathrm{J}_{33} \end{bmatrix} \tag{2.3.24}$$

where

$$\begin{aligned}
\mathrm{J}_{11} &= \lambda^2\mathrm{A}(\mathbf{0}) + \mathrm{E}(\boldsymbol{\Psi}-\boldsymbol{V}) + \mathrm{E}(\boldsymbol{W}-\boldsymbol{\Psi}), \\
\mathrm{J}_{12} &= -\mathrm{E}(\boldsymbol{\Psi}-\boldsymbol{V}). \\
\mathrm{J}_{13} &= -\mathrm{E}(\boldsymbol{W}-\boldsymbol{\Psi}), \\
\mathrm{J}_{21} &= \mathrm{C}\left(\left[\tilde{\boldsymbol{\Psi}}-\tilde{\boldsymbol{V}}\right].\tilde{\boldsymbol{V}}\right) - \rho_v\mathrm{G}(\boldsymbol{\Psi},\boldsymbol{V},\boldsymbol{W}), \\
\mathrm{J}_{22} &= \mathrm{A}\left(\tilde{\boldsymbol{\Psi}}-\tilde{\boldsymbol{V}}\right) - \mathrm{C}\left(\left[\tilde{\boldsymbol{\Psi}}-\tilde{\boldsymbol{V}}\right].\tilde{\boldsymbol{V}}\right) + \rho_v\mathrm{H}(\boldsymbol{\Psi},\boldsymbol{V},\boldsymbol{W}), \\
\mathrm{J}_{23} &= -\rho_v\mathrm{K}(\boldsymbol{\Psi},\boldsymbol{V},\boldsymbol{W}), \\
\mathrm{J}_{31} &= -\mathrm{C}\left(\left[\tilde{\boldsymbol{W}}-\tilde{\boldsymbol{\Psi}}\right].\tilde{\boldsymbol{W}}\right) + \rho_w\mathrm{G}(\boldsymbol{\Psi},\boldsymbol{V},\boldsymbol{W}), \\
\mathrm{J}_{32} &= -\rho_w\mathrm{H}(\boldsymbol{\Psi},\boldsymbol{V},\boldsymbol{W}).
\end{aligned}$$

$$J_{33} = A\left(\tilde{\boldsymbol{W}} - \tilde{\boldsymbol{\Psi}}\right) + C\left(\left[\tilde{\boldsymbol{W}} - \tilde{\boldsymbol{\Psi}}\right], \tilde{\boldsymbol{W}}\right) + \rho_w K(\boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W}).$$

In the above $A(\boldsymbol{B})$, for a vector $\boldsymbol{B}$ in $[\mathcal{N}] \times [\mathcal{N}_D]$, is the matrix $\tilde{A}(\boldsymbol{B})$ minus the columns corresponding to the Dirichlet boundary $\partial\Omega_D$. With $\boldsymbol{\delta}_{pq}$ denoting the Kronecker delta the matrices E, C, G, H and K are defined to be:

$$\begin{aligned} E(\boldsymbol{B})_{pq} &= w_p \delta^2 \exp(B_p) \boldsymbol{\delta}_{pq}, \quad p, q \in \mathcal{N}, \quad \boldsymbol{B} \in [\mathcal{N}], \\ C(\boldsymbol{B}, \boldsymbol{D})_{pq} &= \sum_{l \in \mathcal{N} \cup \mathcal{N}_D} D_l \left(\exp\left(\sum_{r \in \mathcal{N} \cup \mathcal{N}_D} B_r \phi_r \right) \phi_q \nabla \phi_l, \nabla \phi_p \right), \quad p, q \in \mathcal{N}, \\ &\qquad\qquad \boldsymbol{B}, \boldsymbol{D} \in [\mathcal{N}] \times [\mathcal{N}_D], \end{aligned}$$

$$G(\boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W})_{pq} = w_p \sigma \frac{[1 - \exp(W_p - V_p)]\,[\exp(\Psi_p - V_p) - \exp(W_p - \Psi_p)]}{[\exp(W_p - \Psi_p) + \exp(\Psi_p - V_p) + 2]^2} \boldsymbol{\delta}_{pq}, \quad p, q \in \mathcal{N},$$

$$H(\boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W})_{pq} = w_p \sigma \frac{[\exp(\Psi_p - V_p) + \exp(W_p - V_p)\,\{\exp(W_p - \Psi_p) + 2\}]}{[\exp(W_p - \Psi_p) + \exp(\Psi_p - V_p) + 2]^2} \boldsymbol{\delta}_{pq}, \quad p, q \in \mathcal{N},$$

$$K(\boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W})_{pq} = w_p \sigma \frac{[\exp(W_p - \Psi_p) + \exp(W_p - V_p)\,\{\exp(\Psi_p - V_p) + 2\}]}{[\exp(W_p - \Psi_p) + \exp(\Psi_p - V_p) + 2]^2} \boldsymbol{\delta}_{pq}, \quad p, q \in \mathcal{N}.$$

Before proving Newton's Method converges when applied to (2.2.20) we first state Newton's Method formally:

2.3.1 Formal Statement of Newton's Method

Suppose $m \in \mathbb{N}$ and $\mathbf{G} : \mathcal{D} \to \mathbb{R}^m$, where $\mathcal{D}$ is an open subset of $\mathbb{R}^m$. Further, assume there exists an $\boldsymbol{X}^* \in \mathbb{R}^m$ such that $\mathbf{G}(\boldsymbol{X}^*) = \mathbf{0}$. Newton's Method is:

- Guess $\boldsymbol{X}^0 \in \mathbb{R}^m$, an approximation to $\boldsymbol{X}^*$.
- For $k \geq 0$ iterate the following two steps:
  1. Solve: $\mathbf{G}_X(\boldsymbol{X}^k)\boldsymbol{d}^k = -\mathbf{G}(\boldsymbol{X}^k)$ for $\boldsymbol{d}^k$.
  2. Update the solution: $\boldsymbol{X}^{k+1} = \boldsymbol{X}^k + \boldsymbol{d}^k$.

[Here, and in the following, $\mathbf{G}_X(\boldsymbol{X})$ represents the Jacobian of $\mathbf{G}(\boldsymbol{X})$ with respect to $\boldsymbol{X}$].

Further details of Newton's method can be found in, for example [54] and [28].

2.3.2 The Implicit Function Theorem

We use the Implicit Function Theorem to show that Newton's Method converges for a starting guess sufficiently close to solutions with non-singular Jacobian (2.3.24). The result is contained in Corollary 2.3.2, but first we state the Implicit Function Theorem.

Theorem 2.3.1 (The Implicit Function Theorem)
Let $\mathbf{G} : \mathcal{D} \subset \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$, *where* $\mathcal{D}$ *is open. Suppose that there exists constants* ρ_2, γ_1 *and* γ_2, *such that for any* $(\boldsymbol{Y}, \boldsymbol{\zeta}), (\boldsymbol{Y}, \boldsymbol{\eta}), (\boldsymbol{Z}, \boldsymbol{\zeta}) \in \mathcal{D}$:

(A1) $\| \mathbf{G}(\boldsymbol{Y}, \boldsymbol{\zeta}) - \mathbf{G}(\boldsymbol{Y}, \boldsymbol{\eta}) \| \leq \rho_2 \| \boldsymbol{\zeta} - \boldsymbol{\eta} \|$,

(A2) $\| \mathbf{G}_Y(\boldsymbol{Y}, \boldsymbol{\zeta}) - \mathbf{G}_Y(\boldsymbol{Z}, \boldsymbol{\zeta}) \| \leq \gamma_1 \| \boldsymbol{Y} - \boldsymbol{Z} \|$,

(A3) $\| \mathbf{G}_Y(\boldsymbol{Y}, \boldsymbol{\zeta}) - \mathbf{G}_Y(\boldsymbol{Y}, \boldsymbol{\eta}) \| \leq \gamma_2 \| \boldsymbol{\zeta} - \boldsymbol{\eta} \|$

and for $(\boldsymbol{Y}_0, \boldsymbol{\zeta}_0) \in \mathcal{D}$:

(A4) $\mathbf{G}(\boldsymbol{Y}_0, \boldsymbol{\zeta}_0) = \mathbf{0}$,

(A5) $\mathbf{G}_Y(\boldsymbol{Y}_0, \boldsymbol{\zeta}_0)$ *is non-singular.*

Then there exist neighbourhoods $\mathcal{B}(\boldsymbol{\zeta}_0, \varepsilon_\zeta)$, $\mathcal{B}(\boldsymbol{Y}_0, \varepsilon_Y)$ *such that for all* $\boldsymbol{\zeta} \in \mathcal{B}(\boldsymbol{\zeta}_0, \varepsilon_\zeta)$, *there exists* $\boldsymbol{Y}(\boldsymbol{\zeta}) \in \mathcal{B}(\boldsymbol{Y}_0, \varepsilon_Y)$ *with*

(a) $\mathbf{G}(\boldsymbol{Y}(\boldsymbol{\zeta}), \boldsymbol{\zeta}) = \mathbf{0}$,

(b) $\boldsymbol{Y}(\boldsymbol{\zeta})$ *is the unique solution of* $\mathbf{G}(\boldsymbol{Y}, \boldsymbol{\zeta}) = \mathbf{0}$ *in* $\mathcal{B}(\boldsymbol{Y}_0, \varepsilon_Y)$,

(c) $\boldsymbol{Y}(\boldsymbol{\zeta}_0) = \boldsymbol{Y}_0$,

(d) $\mathbf{G}_Y(\boldsymbol{Y}(\boldsymbol{\zeta}), \boldsymbol{\zeta})$ *is non-singular for all* $\boldsymbol{\zeta} \in \mathcal{B}(\boldsymbol{\zeta}_0, \varepsilon_\zeta)$,

(e) $\boldsymbol{Y}(\boldsymbol{\zeta})$ *is continuous with respect to* $\boldsymbol{\zeta} \in \mathcal{B}(\boldsymbol{\zeta}_0, \varepsilon_\zeta)$.

Proof See, for example, [5] or [58] □

The following corollary of the Implicit Function Theorem will be used to prove that the Newton Method given in section 2.3.1 converges.

Corollary 2.3.2 *If (A1)-(A5) of Theorem 2.3.1 applied to*

$$\mathbf{G}(\boldsymbol{Y}, \boldsymbol{\zeta}) = \mathbf{0} \tag{2.3.25}$$

hold, then Newton's Method applied to (2.3.25) converges (locally) to the solution $\boldsymbol{Y}(\boldsymbol{\zeta})$ *for* $\boldsymbol{\zeta}$ *sufficiently close to* $\boldsymbol{\zeta}_0$.

Proof In order to prove that Newton's Method converges to a solution of the system (2.3.25) for a given $\boldsymbol{\zeta}$, we need to show that (see for example [54]):

- (2.3.25) has a solution $\boldsymbol{Y}^*$,
- The Jacobian of (2.3.25) is non-singular at the solution $\boldsymbol{Y}^*$ and
- The Jacobian is Lipschitz continuous in a neighbourhood of the solution $\boldsymbol{Y}^*$.

The first two requirements are conclusions of the Implicit Function Thereom, the third is a hypothesis of the Implicit Function Thereom. Thus if (A1)-(A5) hold for (2.3.25) then Newton's Method will converge. □

2.3.3 Proof of Convergence of Newton's Method

In this section we apply the Implicit Function Theorem to (2.2.20) and prove that the conditions (A1)-(A5) hold. Then Corollary 2.3.2 implies that Newton's method applied to (2.2.20) converges (locally) to a solution, for small enough $\|\boldsymbol{\alpha}\|$.

Here we consider a fixed $\alpha_* > 0$ and the open ball in $[\mathcal{N}] \times [\mathcal{N}_D]$ given by

$$\mathcal{D} := \mathcal{B}(\boldsymbol{X}_0, \alpha_*) \times \mathcal{B}(\boldsymbol{\alpha}_0, \alpha_*) \tag{2.3.26}$$

where

$$\boldsymbol{X}_0 = \left[\boldsymbol{\Psi}_0^T, \mathbf{0}^T, \mathbf{0}^T\right]^T \in [\mathcal{N}]^3, \tag{2.3.27}$$

$$\boldsymbol{\alpha}_0 = \mathbf{0} \in [\mathcal{N}_D] \tag{2.3.28}$$

and $\boldsymbol{\Psi}_0$ solves $\mathbf{F}_1\left(\left[\boldsymbol{\Psi}_0^T, \mathbf{0}^T, \mathbf{0}^T\right]^T, \mathbf{0}\right) = \mathbf{0}$ with $\mathbf{F}_1$ given by (2.2.21)

Before proving that Newton's Method converges we need two lemmas. The first details some properties of the matrices appearing in the discretised semiconductor system and the second shows that $\boldsymbol{\Psi}_0$ is well-defined.

Lemma 2.3.3 *For any vector* $\boldsymbol{B} \in [\mathcal{N}] \times [\mathcal{N}_D]$:

(i) $A(\boldsymbol{B})$ *is an irreducibly diagonally dominant matrix with negative off diagonal elements and strictly positive diagonal elements.*

(ii) *$A(\boldsymbol{B})$ is non-singular and has a strictly positive inverse.*

(iii) *Finally, if D is a positive diagonal matrix, then $A(\boldsymbol{B}) + D$ is also non-singular with a strictly positive inverse.*

Proof In this proof terms from the Section A.2 are used. The reader should refer to this section for the definitions.

The matrix $\tilde{A}(\boldsymbol{B})$ is the stiffness matrix corresponding to the finite element approximation of the operator

$$-\nabla.\left(\exp\left(\sum_{r\in\mathcal{N}\cup\mathcal{N}_D} B_r\phi_r\right)\nabla u(x)\right), \tag{2.3.29}$$

while A($\boldsymbol{B}$) is the matrix $\tilde{A}(\boldsymbol{B})$ minus the columns corresponding to the Dirichlet boundary $\partial\Omega_D$.

The methods of [45] tell us that for a mesh satisfying assumption **(M3)** $\tilde{A}(\boldsymbol{B}) := (a_{p,q})$ has the following properties:

(1) $a_{p,q} \leq 0, \quad p \neq q, \quad p \in \mathcal{N}, \quad q \in \mathcal{N}\cup\mathcal{N}_D.$

(2) $a_{p,p} > 0, \quad p \in \mathcal{N}.$

(3) $a_{p,p} = -\sum_{q\in\mathcal{N}\cup\mathcal{N}_D} a_{p,q}, \quad p \in \mathcal{N}.$

(4) A($\boldsymbol{B}$) is connected.

Properties (1)-(3) imply the required sign condition for A($\boldsymbol{B}$) and that A($\boldsymbol{B}$) is diagonally dominant. Since A($\boldsymbol{B}$) does not include the columns arising from the mesh points on the Dirichlet boundary, (3) also shows that A($\boldsymbol{B}$) has a number of rows which are strictly diagonally dominant (the rows which have non-zero entries corresponding to the mesh points on the Dirichlet boundary of the extended matrix).

A matrix is irreducibly diagonally dominant if it is irreducible, diagonally dominant and has at least one row which is strictly diagonally dominant. It remains to show that A($\boldsymbol{B}$) is irreducible to finish the proof of part (i). However property (4) shows that A($\boldsymbol{B}$) is connected and since a matrix is irreducible if it is connected (Theorem A.2.3) we conclude that A($\boldsymbol{B}$) is irreducible, as required.

It follows from Corollary 1 of Theorem 3.11 of [69] and part (i) of this lemma that the matrix $\mathrm{A}(\boldsymbol{B})$ is non-singular and has a strictly positive inverse, as required for part (ii).

To show part (iii) of the lemma note that adding a positive diagonal matrix to $\mathrm{A}(\boldsymbol{B})$ will not change the sign properties, the connectivity or the diagonal dominance of the matrix. Therefore if we add a positive diagonal matrix to $\mathrm{A}(\boldsymbol{B})$ the resulting matrix will satisfy all the conditions of Corollary 1 of Theorem 3.11 of [69] and applying this corollary completes the proof. □

The following lemma proves that $\boldsymbol{\Psi}_0$ exists and is unique.

Lemma 2.3.4 *There exists a unique* $\boldsymbol{\Psi}_0$ *in* $[\mathcal{N}]$ *such that* $\left[\boldsymbol{\Psi}_0^T, \mathbf{0}^T, \mathbf{0}^T\right]^T$ *solves the problem*

$$\mathbf{F}_1\left(\left[\boldsymbol{\Psi}_0^T, \mathbf{0}^T, \mathbf{0}^T\right]^T, \mathbf{0}\right) = \mathbf{0} \tag{2.3.30}$$

where $\mathbf{F}_1$ *is given by (2.2.21). Furthermore* $\boldsymbol{\Psi}_0$ *is bounded independently of the maximum diameter,* h*, of the triangles in* $\mathcal{T}_h$*.*

Proof The Fréchet derivative of the function given by the left hand side of (2.3.30) evaluated at an arbitrary vector $\boldsymbol{B} \in [\mathcal{N}]$ is $\lambda^2\mathrm{A}(\mathbf{0}) + \mathrm{E}(\boldsymbol{B}) + \mathrm{E}(-\boldsymbol{B})$. Since $\mathrm{E}(\pm\boldsymbol{B})$ is a diagonal matrix with positive entries, it follows from Lemma 2.3.3 that the Fréchet derivative is non-singular and has a positive inverse.

Since the Fréchet derivative has a positive inverse we may apply Theorem 3.3 of [23] to show that there exists a unique finite element solution $\boldsymbol{\Psi}_0$ satisfying (2.3.30), with *a priori* bounds on $\boldsymbol{\Psi}_0$, depending on the Dirichlet boundary data (2.1.9), but independent of the maximum diameter of the triangles in $\mathcal{T}_h$. □

We may now prove the main result:

Theorem 2.3.5 *There exists* ε_α *such that for all* $\boldsymbol{\alpha}$ *in* $[\mathcal{N}_D]$ *with* $\|\boldsymbol{\alpha}\| < \varepsilon_\alpha$ *Newton's method for* $\mathbf{F}(\boldsymbol{X}, \boldsymbol{\alpha}) = \mathbf{0}$ *(given by (2.2.20)) converges to* $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{\alpha})$ *from a starting guess sufficiently close to* $\boldsymbol{X}(\boldsymbol{\alpha})$*.*

Proof The result is obtained from Corollary 2.3.2, which follows from Theorem 2.3.1 with $\boldsymbol{X}_0$ and $\boldsymbol{\alpha}_0$ given by (2.3.27) and (2.3.28).

Thus we now prove that the hypothesis of Theorem 2.3.1 holds for the system given by (2.2.20):

By our choice of $\boldsymbol{X}_0$ and $\boldsymbol{\alpha}_0$ it is certainly true that by Lemma 2.3.4:

$$\mathbf{F}_1(\boldsymbol{X}_0, \boldsymbol{\alpha}_0) = \mathbf{0}$$

and after some calculation:

$$\mathbf{F}_2(\boldsymbol{X}_0, \boldsymbol{\alpha}_0) = \mathbf{F}_3(\boldsymbol{X}_0, \boldsymbol{\alpha}_0) = \mathbf{0}.$$

Thus assumption (A4) of Theorem 2.3.1 holds.

To verify assumption (A5), we note that

$$\mathrm{J}(\boldsymbol{X}_0, \boldsymbol{\alpha}_0) = \begin{bmatrix} \mathrm{J}^0_{11} & \mathrm{J}^0_{12} & \mathrm{J}^0_{13} \\ 0 & \mathrm{J}^0_{22} & \mathrm{J}^0_{23} \\ 0 & \mathrm{J}^0_{32} & \mathrm{J}^0_{33} \end{bmatrix}, \tag{2.3.31}$$

where

$$\begin{aligned}
\mathrm{J}^0_{11} &= \lambda^2 \mathrm{A}(\mathbf{0}) + \mathrm{E}(\boldsymbol{\Psi}_0) + \mathrm{E}(-\boldsymbol{\Psi}_0), \\
\mathrm{J}^0_{12} &= -\mathrm{E}(\boldsymbol{\Psi}_0), \\
\mathrm{J}^0_{13} &= -\mathrm{E}(-\boldsymbol{\Psi}_0), \\
\mathrm{J}^0_{22} &= \mathrm{A}\left(\left[\boldsymbol{\Psi}_0^T, \mathbf{0}^T\right]^T\right) + \rho_v \mathrm{H}_0(\boldsymbol{\Psi}_0), \\
\mathrm{J}^0_{23} &= -\rho_v \mathrm{H}_0(\boldsymbol{\Psi}_0), \\
\mathrm{J}^0_{32} &= -\rho_w \mathrm{H}_0(\boldsymbol{\Psi}_0), \\
\mathrm{J}^0_{33} &= \mathrm{A}\left(-\left[\boldsymbol{\Psi}_0^T, \mathbf{0}^T\right]^T\right) + \rho_w \mathrm{H}_0(\boldsymbol{\Psi}_0).
\end{aligned}$$

In addition to those matrices already given, H_0 is defined by:

$$\mathrm{H}_0(\boldsymbol{\Psi}_0)_{pq} = w_p \frac{\sigma}{\left[2\cosh\left((\boldsymbol{\Psi}_0)_p\right) + 2\right]} \delta_{pq}, \quad p, q \in \mathcal{N}.$$

To prove (A5) of the Implicit Function Theorem we aim to show that $\mathrm{J}(\boldsymbol{X}_0, \boldsymbol{\alpha}_0)$, given by (2.3.31), is essentially diagonally dominant (Definition A.2.5). If we can show this then it follows from Theorem 6.4.10 of [37] that $\mathrm{J}(\boldsymbol{X}_0, \boldsymbol{\alpha}_0)$ is non-singular.

The matrix is essentially diagonally dominant if it is diagonally dominant with a number of rows that are strictly diagonally dominant. To show this consider the rows

in the first block of (2.3.31) we note that exactly what has been added to the diagonal of A($\mathbf{0}$) has been subtracted off from other elements in the same row. Since A($\mathbf{0}$) is diagonally dominant it must follow that all rows in the first block row of (2.3.31) are also diagonally dominant. Further we have seen in Lemma 2.3.3 that a number of rows of A($\mathbf{0}$) are strictly diagonally dominant, implying the same must be true of the rows in the first block of (2.3.31). The same conclusion can be reached for the rows in the second and third blocks of (2.3.31).

To show J$(\boldsymbol{X}_0, \boldsymbol{\alpha}_0)$ is essentially diagonally dominant we must show that, for each node γ of the matrix (here the term node is understood in the sense of graph theory: Appendix A), there is at least one node, node μ say, such that node γ is connected to node μ and row μ is strictly diagonally dominant. Since A($\boldsymbol{B}$) is connected, for every $\boldsymbol{B}$, it is easy to see that the nodes associated with the rows in the first block of (2.3.31) are connected to each other and since at least one of these rows is strictly diagonally dominant, the required condition is satisfied for these nodes. The same conclusion can be reached for all rows in the second and third block rows. Proving that J$(\boldsymbol{X}_0, \boldsymbol{\alpha}_0)$ is essentially diagonally dominant.

In conclusion we may appeal to Theorem 6.4.10 of [37] to show that (2.3.31) is non-singular, proving (A5).

Finally we need to show (A1)-(A3) of the Implicit Function Theorem:

Since $\mathbf{F}(\boldsymbol{X}, \boldsymbol{\alpha})$ is continuously differentiable in $\boldsymbol{X}$ and $\boldsymbol{\alpha}$ and since each derivative is bounded when $(\boldsymbol{X}, \boldsymbol{\alpha})$ lies in the bounded set $\mathcal{D}$, we may use the mean value theorem to show (A1).

Similarly $\mathbf{F}_X(\boldsymbol{X}, \boldsymbol{\alpha})$ given by (2.3.24) is continuously differentiable in $\boldsymbol{X}$ and $\boldsymbol{\alpha}$, thus if $(\boldsymbol{X}, \boldsymbol{\alpha})$ are members of $\mathcal{D}$, (A2) and (A3) follow by the mean value theorem.

Thus the Implicit Function Theorem holds for the system given by (2.2.20) and so by Corollary 2.3.2, Theorem 2.3.5 is proved. □

2.4 Numerical Results for Newton's Method Applied to the PN Diode

2.4.1 Newton's Method

In this section Newton's method is applied to the finite element discretisation of the system modelling a PN diode in one dimension. Results show that Newton's method applied to the system only converges for sufficiently small applied voltage.

The finite element method with mass lumping is applied to the semiconductor system (2.1.1)-(2.1.3) in one dimension with $\Omega = [0, 1]$. Since the aim is to model a PN diode the doping profile chosen is -1 on $[0, 1/2)$ and +1 on $(1/2, 1]$. The equations are discretised with respect to a uniform grid with n interior mesh points. Newton's method, as described in Section 2.3.1, is applied to the finite element system. The linear systems produced by Newton's method are solved using the block Gauss-Seidel method, where the blocking is with respect to variable (Ψ, V and W). This procedure is frequently referred to as a "Newton Gauss-Seidel" method in the literature. The initial guess for ψ, v and w is the doping profile scaled to match the relevant Dirichlet boundary conditions. The results are contained in Table 2.1.

Number of interior mesh points	Voltage applied at 0	Voltage applied at 1	Number of Newton iterations
9	0	0.1	6
9	0	0.15	11
9	0	0.2	Diverges
9	0.1	0	Diverges
9	7.3	7.45	11
21	0	0.1	7
21	0	0.2	Diverges

Table 2.1: Newton's method applied to the one dimensional discretisation of a PN diode on a uniform grid. The iteration is stopped when the change in $\left(\boldsymbol{\Psi}^T, \boldsymbol{V}^T, \boldsymbol{W}^T\right)^T$ is less than 5×10^{-5}.

Comparing the results in Table 2.1 with those for Gummel's method discussed in Chapter 3 we see that both methods converge for small reverse bias (the contact at the n-type region has a more positive voltage applied than the contact at the p-type region, as discussed in Chapter 1). Newton's method converges for a smaller range of applied voltages and will not converge in the forward bias situation. It is generally asserted in

the literature that Gummel's method is more robust than Newton's method, however Gummel's method converges linearly while Newton's method converges quadratically when sufficiently close to the true solution. Nevertheless both methods break down as the voltage is increased unless a better starting guess can be found.

One way of overcoming the problem of the poor quality starting guesses is to use Newton's method together with a continuation procedure:

2.4.2 Newton's Method with Continuation

In this section we combine a continuation scheme with Newton's method. We aim to try to find the solution to the semiconductor equations for any applied voltage by solving a series of problems for intermediate voltages. The starting guess for the problem with slightly increased voltage is based on the previous solution, rather than the scaling of the doping profile (used in the previous section).

As a way of introducing the method consider a device with two contacts, assume we have a (scaled) applied voltage of α_0 at the left hand contact and a (scaled) applied voltage of α_1 at the right hand contact. Then the nonlinear system can be written in the following way: Seek $\boldsymbol{X}$ such that:

$$\mathbf{F}(\boldsymbol{X}, \alpha_0, \alpha_1) = \mathbf{0}. \tag{2.4.32}$$

The continuation method seeks a solution to the problem

$$\mathbf{F}(\boldsymbol{X}, \alpha_0, \alpha_0 + k(\alpha_1 - \alpha_0)) = \mathbf{0} \tag{2.4.33}$$

for k between 0 and 1. The aim is to find a solution when $k = 1$. We assume we know, or can easily find, the solution when $k = 0$. Rewrite (2.4.33), for convenience, as:

$$\mathbf{F}(\boldsymbol{X}, k) = \mathbf{0}. \tag{2.4.34}$$

Assume we have a solution $\boldsymbol{X}_i$ to the system:

$$\mathbf{F}(\boldsymbol{X}_i, k_i) = \mathbf{0}, \quad k_i \in [0, 1).$$

We try to solve

$$\mathbf{F}(\boldsymbol{X}, k_{i+1}) = \mathbf{0}, \quad \text{where } k_{i+1} = k_i + \Delta k_i. \tag{2.4.35}$$

To do this we follow standard procedure and differentiate (2.4.34) with respect to k to obtain

$$\mathrm{J}(\boldsymbol{X}, k)\frac{\partial \boldsymbol{X}}{\partial k} + \mathbf{F}_k(\boldsymbol{X}, k) = \mathbf{0}, \tag{2.4.36}$$

where $\mathrm{J}(\boldsymbol{X}, k)$ represents the Jacobian of $\mathbf{F}$ with respect to $\boldsymbol{X}$. Rearranging (2.4.36) we have

$$\frac{\partial \boldsymbol{X}}{\partial k} = -\mathrm{J}(\boldsymbol{X}, k)^{-1}\mathbf{F}_k(\boldsymbol{X}, k).$$

Using one step of Euler's method when $k = k_i$:

$$\frac{\boldsymbol{X}_E - \boldsymbol{X}_i}{\Delta k_i} = -\mathrm{J}(\boldsymbol{X}_i, k_i)^{-1}\mathbf{F}_k(\boldsymbol{X}_i, k_i)$$

or

$$\boldsymbol{X}_E = \boldsymbol{X}_i - \Delta k_i \mathrm{J}(\boldsymbol{X}_i, k_i)^{-1}\mathbf{F}_k(\boldsymbol{X}_i, k_i). \tag{2.4.37}$$

$\boldsymbol{X}_E$ will be the initial guess in an application of Newton's method for solving (2.4.35). If the iterations start to diverge we reduce the step size, Δk_i, and start again from the previous converged solution $\boldsymbol{X}_i$.

Implementation

We have applied these ideas to the drift-diffusion semiconductor system (2.1.1)-(2.1.3) modelling the PN diode on the one dimensional domain $\Omega = [0, 1]$. We seek a finite element solution for (scaled) applied voltages α_0 at $x = 0$ and α_1 at $x = 1$. For each $k_i \in [0, 1]$ we seek the finite element solutions Ψ_i, V_i and W_i satisfying the mass lumped finite element system and the boundary conditions:

$$\begin{aligned} \Psi_i(0) &= -\beta + \alpha_0, \quad V_i(0) = W_i(0) = \alpha_0, \\ \Psi_i(1) &= -\beta + \alpha_0 + k_i(\alpha_1 - \alpha_0), \quad V_i(0) = W_i(0) = \alpha_0 + k_i(\alpha_1 - \alpha_0). \end{aligned}$$

In the above $\beta = \sinh^{-1}(1/2\delta^2)$ is the intrinsic voltage of the device. As before, define $\boldsymbol{\Psi}_i, \boldsymbol{V}_i$ and $\boldsymbol{W}_i$ to be the vector of values of Ψ_i, V_i and W_i at the interior mesh points of Ω. The extended vector $\tilde{\boldsymbol{\Psi}}_i$ is defined to be $[\Psi_i(0), \boldsymbol{\Psi}_i^T, \Psi_i(1)]^T$. $\tilde{\boldsymbol{V}}_i$ and $\tilde{\boldsymbol{W}}_i$ are

defined analogously. Defining $\boldsymbol{X}_i = \left[\boldsymbol{\Psi}_i^T, \boldsymbol{V}_i^T, \boldsymbol{W}_i^T\right]^T$ then let $\mathbf{F}(\boldsymbol{X}_i, k_i)$ be the usual mass lumped finite element discretisation of the semiconductor equations at $\boldsymbol{X}_i$.

To use the continuation step, (2.4.37), to obtain an initial guess for the finite element solution for a system with slightly increased voltage we need the derivative of $\mathbf{F}$ with respect to k and the Jacobian of $\mathbf{F}$. $\mathbf{F}_k$ is easily seen to be:

$$\mathbf{F}_k(\boldsymbol{X}_i, k_i) = \begin{bmatrix} \lambda^2(\alpha_1 - \alpha_0)\tilde{A}(\mathbf{0}) \begin{bmatrix} 0 \\ \mathbf{0} \\ 1 \end{bmatrix} \\ (\alpha_1 - \alpha_0)\tilde{A}\left(\tilde{\boldsymbol{\Psi}}_i - \tilde{\boldsymbol{V}}_i\right) \begin{bmatrix} 0 \\ \mathbf{0} \\ 1 \end{bmatrix} \\ (\alpha_1 - \alpha_0)\tilde{A}\left(\tilde{\boldsymbol{W}}_i - \tilde{\boldsymbol{\Psi}}_i\right) \begin{bmatrix} 0 \\ \mathbf{0} \\ 1 \end{bmatrix} \end{bmatrix}.$$

Since (2.4.37) is used to obtain an initial guess for the solution with increased voltage full accuracy is not needed, therefore we do not use the full Jacobian in (2.4.37), instead we replace $\mathrm{J}(\boldsymbol{X}, k)$ by the simpler approximation $\hat{\mathrm{J}}(\boldsymbol{X}, k)$ defined by:

$$\hat{\mathrm{J}}(\boldsymbol{X}_i, k_i) = \begin{bmatrix} \lambda^2\mathrm{A}(\mathbf{0}) + \mathrm{E}(\boldsymbol{\Psi}_i - \boldsymbol{V}_i) & -\mathrm{E}(\boldsymbol{\Psi}_i - \boldsymbol{V}_i) & -\mathrm{E}(\boldsymbol{W}_i - \boldsymbol{\Psi}_i) \\ \quad +\mathrm{E}(\boldsymbol{W}_i - \boldsymbol{\Psi}_i) & & \\ 0 & \mathrm{A}\left(\tilde{\boldsymbol{W}}_i - \tilde{\boldsymbol{\Psi}}_i\right) & -\rho_v\mathrm{K}(\boldsymbol{\Psi}_i, \boldsymbol{V}_i, \boldsymbol{W}_i) \\ & +\rho_v\mathrm{H}(\boldsymbol{\Psi}_i, \boldsymbol{V}_i, \boldsymbol{W}_i) & \\ 0 & -\rho_w\mathrm{H}(\boldsymbol{\Psi}_i, \boldsymbol{V}_i, \boldsymbol{W}_i) & \mathrm{A}\left(\tilde{\boldsymbol{W}}_i - \tilde{\boldsymbol{\Psi}}_i\right) \\ & & +\rho_w\mathrm{K}(\boldsymbol{\Psi}_i, \boldsymbol{V}_i, \boldsymbol{W}_i) \end{bmatrix}.$$

To start the method off we take $k_0 = 0$ and the first stage of the continuation process collapses to finding Ψ_0 such that the extended vector $\tilde{\boldsymbol{\Psi}}_0$ satisfies:

$$\lambda^2\tilde{A}(\mathbf{0})\tilde{\boldsymbol{\Psi}}_0 + \boldsymbol{e}(\boldsymbol{\Psi}_0) - \boldsymbol{e}(-\boldsymbol{\Psi}_0) - \boldsymbol{d}. \tag{2.4.38}$$

where $\Psi_0(0) = -\beta + \alpha_0$, $\Psi_0(1) = \beta + \alpha_0$. (2.4.38) can easily be solved by Newton's Method with an initial guess based on the scaled doping profile (scaled to fit the boundary conditions). V_0 and W_0 are known to be identically equal to α_0.

We implement our continuation procedure with a step size of $\Delta k_i = 0.1$, for all i. If the iteration starts to diverge we backtrack by halving Δk_i and restarting from the previous converged solution, $\boldsymbol{X}_i$.

Results for the continuation procedure described above are presented in Table 2.2. The left hand applied voltage α_0 is always taken to be zero for these results. As we initially take $k_0 = 0$ and $\Delta k_i = 0.1$ the minimum number of continuations steps required is 11, more than 11 continuation steps indicate that the program has backtracked at some point.

The results in Table 2.2 show that the strategy works well for reasonably large applied voltages (only 15 continuation steps are needed for an applied voltage of 0.5 volts). We have tried to be optimistic in our approach and have taken fixed Δk_i, this pays off for the initial continuation steps at least, but does mean backtracking becomes necessary as k_i becomes large.

Number of interior mesh points	Voltage applied at 1	Number of continuation steps	Number of Newton iterations
9	0.1	11	31
9	0.2	11	41
9	0.4	11	61
9	0.5	15	79
9	1.0	54	202
21	0.2	11	31
21	0.5	24	105
21	1.0	65	202
21	1.2	102	302

Table 2.2: Newton's method with continuation applied to the one dimensional discretisation of a PN diode on a uniform grid. The voltage applied at the left hand boundary is always zero. The iteration is stopped when the change in $\left(\boldsymbol{\Psi}^T, \boldsymbol{V}^T, \boldsymbol{W}^T\right)^T$ is less than 5×10^{-5}.

Chapter 3

The Alternative Nodal Factorisation Method

3.1 Introduction

In this chapter we analyse the convergence of a non-standard method for solving algebraic equations arising from finite element discretisations of nonlinear elliptic systems.

In standard methods for solving such systems, the equations are ordered in the natural way inherited from the ordering of the PDEs themselves and the Jacobian arising in Newton's method then inherits a blocking from this ordering. For example if there are 3 PDEs discretised on a mesh with ν degrees of freedom, then the Jacobian will have a 3×3 block structure with each block of size $\nu \times \nu$. (For convenience it is assumed that all the PDEs are discretised on the same mesh). A typical approximate Newton scheme for this system may be obtained by applying some block Jacobi or block Gauss-Seidel iteration based on this blocking. This leads to successive solutions of each individual PDE in turn with the coupling between the PDEs neglected. Such a "Newton Jacobi" or "Newton Gauss-Seidel" scheme is typical in device modelling, but works well only if the coupling between the PDEs themselves is weak in comparison to the coupling between the values of the solution of a single PDE at different mesh points of the mesh, for the latter coupling is preserved in the iterative method, while the former is broken. However in some applications the latter coupling is the weaker of the two. In particular, in the highly nonlinear systems arising in semiconductor modelling, especially in the presence of high currents, this has been found to be the case [8].

For such applications it is then natural to order the equations differently and to group the unknowns corresponding to each mesh point together. This we call the **Alternative Nodal Factorisation** (ANF). Using this reordering the Jacobian then has a $\nu \times \nu$ block structure and a Jacobi or Gauss-Seidel block iteration method which solves 3×3 systems for updates to the three unknowns at each mesh point can be written down. This preserves the coupling between the different unknowns at a single mesh point, but breaks the coupling betwecn solution values at different mesh points. This re-blocking has been found to be helpful in device modelling in some circumstances [33].

Since only ν local 3×3 systems have to be resolved at each (outer) iteration, instead of 3 global $\nu \times \nu$ systems, the cost per iterate is low. However it is expected that the rate of convergence of such a scheme will deteriorate as $\nu \to \infty$ in the same way as Jacobi or Gauss-Seidel deteriorates. It would then be natural to consider versions which solve for all the variables at several nearby mesh points simultaneously and (possibly) with the addition of a coarse mesh correction such as is used in one step of the corresponding nonlinear multi-grid method.

To date there is no rigorous theory for this type of iteration. Even the Jacobi-ANF method has so far been justified only with empirical evidence, [8], [30] and [33]. In this chapter the convergence of this simplest version of the ANF method for the full semiconductor system in one dimension is proved. This result is proved with the aid of the recent theory of Dryja and Hackbusch, [29], which shows that if a certain type of linearised subspace iteration converges, then the corresponding nonlinear version also converges.

To apply the theory of [29] to the semiconductor equations we proceed as follows: Using graph theory we show that the Jacobi-ANF iteration applied to the linearisation of the discretised semiconductor equations converges when there is no applied voltage across the device. A perturbation argument extends the convergence of the linearised iteration to the case of small applied voltage. An application of [29] then shows that the nonlinear Jacobi-ANF method converges in the case of small applied voltage. Numerical results show that the method does indeed work and, when compared to standard numerical methods for solving the semiconductor equations extends the range of applied voltages it is possible to solve for.

In this chapter we also discuss Gummel's method - one of the most common methods

used in semiconductor device modelling. Gummel's method is a group of nonlinear block Gauss-Seidel iterations where the blocking is with respect to PDE. The method neglects the coupling between the PDEs. Results show that this method converges for a smaller range of applied voltages than the Jacobi-ANF method, supporting our view that it is important to preserve the coupling between the PDEs in the iterative method.

3.2 The Method

In this section we introduce the Jacobi-ANF method in a general context. From now on we just refer to this as "the ANF iteration", although there are other versions as explained in the previous section.

Consider solving a system of (generally nonlinear) elliptic partial differential equations on some domain together with appropriate boundary conditions. The system is written in terms of scalar PDEs as follows:

$$L(z) = \begin{pmatrix} L_1(z_1, z_2, \ldots, z_m) \\ L_2(z_1, z_2, \ldots, z_m) \\ \vdots \\ L_m(z_1, z_2, \ldots, z_m) \end{pmatrix} = 0 \tag{3.2.1}$$

where $z(\boldsymbol{x}) = (z_1(\boldsymbol{x}), z_2(\boldsymbol{x}), \ldots, z_m(\boldsymbol{x}))^T \in \mathbb{R}^m$ is a vector valued function containing the m solutions of the m PDEs. Applying the finite element method to this system results in the new set of equations:

$$L_h(\boldsymbol{z}) = \begin{pmatrix} L_{1,h}(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m) \\ L_{2,h}(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m) \\ \vdots \\ L_{m,h}(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m) \end{pmatrix} = 0. \tag{3.2.2}$$

If there are ν degrees of freedom associated with the finite element method, then each $\boldsymbol{z}_i$ is a vector with ν entries (the values of $\boldsymbol{z}_i$ at these degrees of freedom) and the system represents $m\nu$ equations in $m\nu$ unknowns.

The alternative nodal factorisation (ANF) method is an iterative method which updates the vector $(\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \ldots, \boldsymbol{z}_m^T)^T$ at each degree of freedom in turn, until convergence. The method can be expressed as:

The ANF Method

1 Make an initial guess, $z^0 = \left((z_1^0)^T, (z_2^0)^T, \ldots, (z_m^0)^T\right)^T$, to the solution of the system (3.2.2). Set $k = 0$.

2 For $j = 1, 2, \ldots, \nu$, find

$$\left(\left(z_1^{k+1}\right)_j, \left(z_2^{k+1}\right)_j, \ldots, \left(z_m^{k+1}\right)_j\right)^T,$$

such that the vectors:

$$\left(\tilde{z}_i^{k+1,j}\right)_l = \begin{cases} \left(z_i^k\right)_l & l \neq j \\ \left(z_i^{k+1}\right)_l & l = j \end{cases} \quad l = 1, 2, \ldots, v, \; i = 1, 2, \ldots, m,$$

satisfy:

$$\begin{pmatrix} L_{1,h}(\tilde{z}_1^{k+1,j}, \tilde{z}_2^{k+1,j}, \ldots, \tilde{z}_m^{k+1,j}) \\ L_{2,h}(\tilde{z}_1^{k+1,j}, \tilde{z}_2^{k+1,j}, \ldots, \tilde{z}_m^{k+1,j}) \\ \vdots \\ L_{m,h}(\tilde{z}_1^{k+1,j}, \tilde{z}_2^{k+1,j}, \ldots, \tilde{z}_m^{k+1,j}) \end{pmatrix} = 0. \tag{3.2.3}$$

3 Set $z^{k+1} = \left(\left(z_1^{k+1}\right)^T, \left(z_2^{k+1}\right)^T, \ldots, \left(z_m^{k+1}\right)^T\right)^T$ to be the vector whose values were calculated in 2.

4 If the norm difference between z^k and z^{k+1} is less than the required tolerance, then stop. Otherwise set $k = k + 1$ and return to step 2.

Solving (3.2.3) corresponds to seeking the m unknowns associated with the jth mesh point while holding all the other unknowns fixed at the values calculated in the previous outer iteration. This is done for each $j = 1, 2, \ldots, \nu$ and may be implemented in parallel. Although we may have to repeat steps 2-4 many times for convergence, solving (3.2.3) should be relatively easy for small m. The cost of step 2 is $O(\nu m^3)$ which is to be compared with $O(\nu^3 m)$ for a Jacobi iterate based on standard blocking.

This algorithm can be thought of as a nonlinear block Jacobi iteration, where the unknown variables at each mesh point are grouped together. This idea is further explored in the next section.

3.2.1 Example: The ANF Method Applied to a Linear System

If $L(z) = b$ is a system of m linear differential equations then a typical discretisation of this system can be written in the form:

$$L_h(\boldsymbol{z}) = A\boldsymbol{z} = \boldsymbol{b}, \tag{3.2.4}$$

where

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,m} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,m} \end{bmatrix},$$

$$\boldsymbol{z} = \left(\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \ldots, \boldsymbol{z}_m^T\right)^T \text{ and } \boldsymbol{b} = \left(\boldsymbol{b}_1^T, \boldsymbol{b}_2^T, \ldots, \boldsymbol{b}_m^T\right)^T$$

Each $A_{i,j}$ represents the discretisation of the differential operator in the ith equation which operates on the jth variable $\boldsymbol{z}_j$. If the discretisation is obtained on a mesh with ν degrees of freedom, for example, then each $A_{i,j}$ is of size $\nu \times \nu$. Applying the ANF algorithm (introduced in the previous section) to (3.2.4) simply yields the block Jacobi method for the reblocked system:

$$\tilde{A}\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{b}} \tag{3.2.5}$$

where the blocking is with respect to individual points in the mesh, i.e. $\tilde{A}$ takes the form:

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{1,1} & \tilde{A}_{1,2} & \cdots & \tilde{A}_{1,v} \\ \tilde{A}_{2,1} & \tilde{A}_{2,2} & \cdots & \tilde{A}_{2,v} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{A}_{v,1} & \tilde{A}_{v,2} & \cdots & \tilde{A}_{v,v} \end{bmatrix}, \tag{3.2.6}$$

$$\tilde{\boldsymbol{z}} = \left([(\boldsymbol{z}_1)_1, (\boldsymbol{z}_2)_1, \ldots, (\boldsymbol{z}_m)_1], \ldots, [(\boldsymbol{z}_1)_v, \ldots, (\boldsymbol{z}_m)_v]\right)^T$$

and $\tilde{\boldsymbol{b}}$ is defined in an analogous way. Each of the individual blocks $\tilde{A}_{i,j}$ is of size $m \times m$ and represents the coupling between the unknowns $((\boldsymbol{z}_1)_i, (\boldsymbol{z}_2)_i, \ldots, (\boldsymbol{z}_m)_i)$ and $((\boldsymbol{z}_1)_j, (\boldsymbol{z}_2)_j, \ldots, (\boldsymbol{z}_m)_j)$, i.e. the collection of nodal values of the m differential equations at the ith and jth mesh points respectively.

A and $\tilde{A}$ are related via a permutation matrix P: $\tilde{A} = PAP^T$. In fact:

$$\tilde{A}_{i,j} = \begin{bmatrix} (A_{1,1})_{i,j} & (A_{1,2})_{i,j} & \cdots & (A_{1,m})_{i,j} \\ (A_{2,1})_{i,j} & (A_{2,2})_{i,j} & \cdots & (A_{2,m})_{i,j} \\ \vdots & \vdots & \ddots & \vdots \\ (A_{m,1})_{i,j} & (A_{m,2})_{i,j} & \cdots & (A_{m,m})_{i,j} \end{bmatrix}.$$

The structure of $\tilde{A}_{i,j}$ highlights some important features of the ANF method. Firstly the sparsity pattern of the block matrix $\tilde{A}$ represents the connectivity of the mesh while the individual blocks reflect the coupling between the partial differential equations. Since the ANF method consists of a block Jacobi iteration applied to $\tilde{A}$ it is reasonable to expect it to work better than the standard method when the variables at a single mesh point are strongly coupled together.

See [8] for further discussion on $\tilde{A}$ and its relationship to A.

Several authors have used the linear and nonlinear form of the ANF method. In [33] the linear form is used as a preconditioner for a Newton-Krylov approach to solving the drift-diffusion semiconductor equations and turns out to be one of the fastest methods tested.

In [30] the nonlinear ANF method is applied to the one dimensional time dependent field phase equations: Find θ and u such that:

$$\begin{aligned} c\theta_t + \frac{l}{2}u_t &= k\Delta\theta + f, \\ \tau u_t = \gamma\Delta u &- \psi'(u) + \alpha\theta \end{aligned}$$

where ψ is a double well potential, f is a volumetric heat source or sink and all other parameters are considered to be constant. Elliott and Gardiner prove that under certain restrictions the ANF method applied to a linear system related to the field phase equations converges. The proof relates the convergence of Jacobi's method applied to the finite element discretisation of Poisson's equation (which is known to converge due to the diagonal dominance of the discretised Laplacian operator) to the convergence of the ANF method applied to the linear system. However the proof uses Fourier analysis and this is restricted to uniform meshes and certain types of boundary conditions. Such restrictions are inappropriate in the context of semiconductor equations.

3.3 Application of the ANF Method to the Semiconductor System

In this section the convergence of the ANF method applied to the nonlinear drift-diffusion semiconductor equations is considered. It is proved that the method converges for small applied voltage. As a comparison to the ANF method we introduce later in this section Gummel's method (a type of nonlinear Gauss-Seidel iteration commonly used in device modelling) and give numerical results for this method. It will be shown at the end of this chapter that the ANF method applied to a PN diode problem converges for a large range of voltages in reverse bias and also converges for small forward bias.

In order to prove the convergence of the nonlinear iteration the theory of Dryja and Hackbusch, [29], is used. This paper is concerned with subspace iterations, in which each step consists of an approximate solution of the problem in appropriate subspaces of the full solution space. [29] proves that, under certain conditions, a subspace iteration method for a given nonlinear problem converges, providing it also converges for the linearised problem about the true solution of the nonlinear problem. The nonlinear and linearised ANF method can be expressed as a subspace iteration in much the same way as the block Jacobi iteration.

To prove that the ANF method applied to the nonlinear drift-diffusion semiconductor equations converges we first show, using graph theory, that the ANF iteration applied to the linearisation of the discretised semiconductor equations converges when there is no applied voltage across the device. A perturbation argument (with respect to voltage) extends the proof to the case of small applied voltage. Finally, an application of [29] completes the proof of convergence for small applied voltage.

The Dryja-Hackbusch theory is discussed in the next section, while the necessary graph theory is discussed in Appendix A.

3.3.1 Dryja-Hackbusch Theory

Dryja and Hackbusch consider in [29] finite dimensional nonlinear problems of the following form: Find $x \in D \subset X$ such that:

$$F(x) = 0. \tag{3.3.7}$$

where X is a finite dimensional space with norm $\|\cdot\|$. The problem (3.3.7) may, for example, be a finite element discretisation of a nonlinear boundary value problem. Dryja and Hackbusch make the following assumptions on (3.3.7):

H1 There exists a solution $x^* \in \mathcal{D}$ to (3.3.7).

H2 There exists a neighbourhood, $U \subset \mathcal{D}$, of x^*, such that x^* is the locally unique solution of (3.3.7) in U.

H3 The Fréchet derivative, $F_x(x)$, of $F(x)$ exists at x^* and is non-singular.

H4 There exists a uniformly bounded linear operator $DF(x', x'') \in L(X, X)$, defined for all $x', x'' \in X$, such that

- $F(x') - F(x'') = DF(x', x'')(x' - x'')$ and
- $\|DF(x', x'') - F_x(x^*)\| \to 0$ as $x', x'' \to x^*$, where $\|\cdot\|$ is the operator norm on $L(X, X)$ induced by the norm $\|\cdot\|$ on X.

Remark 3.3.1 *Under appropriate assumptions it can be shown that* **[H2]** *and* **[H4]** *follow from* **[H1]** *and* **[H3]**, *we state the assumptions in the above form to be consistent with [29].*

The paper [29] is concerned with iteratively solving (3.3.7) using a subspace iteration method. This includes, as a special case, various domain decomposition methods - nonlinear versions of those found in [19].

The subspace iteration method in [29] is determined by choosing a set of disjoint spaces, X_κ ($\kappa \in I$, where I is a finite index set), together with linear injective mappings

$$p_\kappa : X_\kappa \to X,$$

providing prolongation operators and linear surjective mappings

$$r_\kappa : X \to X_\kappa$$

which provide restriction operators. It is usual to take r_κ to be the adjoint of p_κ with respect to the Euclidean inner product, i.e. $r_\kappa = p_\kappa^T$.

It is also required that the subspaces $p_\kappa X_\kappa$ cover the whole of X i.e. that

$$X = \sum_{\kappa \in I} p_\kappa X_\kappa,$$

in which case the true solution x of (3.3.7) can be expressed as $x = \sum_{\kappa \in I} p_\kappa x_\kappa$ with $x_\kappa \in X_\kappa$.

In order to prove that the nonlinear subspace iteration method introduced below converges, Dryja and Hackbusch make one further assumption on the problem:

H5 $r_\kappa F_x(x^*) p_\kappa : X_\kappa \to X_\kappa$ is invertible for each $\kappa \in I$.

Roughly this means that the Jacobian of (3.3.7) should be invertible in each of the subspaces X_κ.

With these assumptions we can now define the general method.

The nonlinear subspace iteration

Given an approximation, $\tilde{x}$, to the solution, x^*, of (3.3.7) in the neighbourhood U, one step of the nonlinear subspace iteration consists of seeking, for each $\kappa \in I$, a $\delta_\kappa \in X_\kappa$ such that

$$r_\kappa F(\tilde{x} - p_\kappa \delta_\kappa) = 0. \tag{3.3.8}$$

The new estimate, $\hat{x}$, of x^* is then given by:

$$\hat{x} = \tilde{x} - \omega \sum_{\kappa \in I} p_\kappa \delta_\kappa \tag{3.3.9}$$

where ω is a damping parameter to be determined. (3.3.8) and (3.3.9) are repeated until convergence.

This general scheme includes a wide variety of familiar special cases. For example if (3.3.7) represents the discretisation of a single PDE then X consists of the space of vectors defined at the free mesh points of the discretisation. The X_κ may denote, for example, vectors with support within small subdomains of the domain of the PDE ("local spaces") or globally defined vectors on suitable coarsening of the original mesh ("coarse spaces").

A simple example arises when (3.3.7) is a system of linear equations in $\mathbb{R}^n$ and the X_κ are the standard basis vectors in $\mathbb{R}^n$, then (3.3.9) is simply the damped Jacobi

method. The ANF method introduced in Section 3.2 is a somewhat more sophisticated example of this general scheme as we shall see in Section 3.3.2.

Convergence of the nonlinear subspace iteration

The proof of the convergence of the nonlinear subspace iteration given in [29] depends on the assumption that the (linear) subspace iteration applied to the linearisation of the nonlinear problem (3.3.7) at the true solution, x^*, converges. In other words, the linear problem considered is:

Find x such that:

$$Ax = b, \tag{3.3.10}$$

where $A := F_x(x^*)$ and $b := Ax^*$. Applying the linear subspace iteration to this problem results in the iterative method:

Make an initial guess, x^0, to the solution of (3.3.10). For $l = 0, 1, \ldots$, find

$$x^{l+1} = x^l - \omega \sum_{\kappa \in I} p_\kappa A_\kappa^{-1} r_\kappa (Ax^l - b) \tag{3.3.11}$$

where

$$A_\kappa = r_\kappa A p_\kappa \tag{3.3.12}$$

for each $\kappa \in I$.

This linear iteration has iteration matrix

$$M_\omega := I - \omega \sum_{\kappa \in I} p_\kappa A_\kappa^{-1} r_\kappa A.$$

It is well know that the iteration converges provided ω, the spaces and the mappings are chosen such that:

$$\|M_\omega\| \leq \sigma_\omega < 1. \tag{3.3.13}$$

In fact it turns out that if (3.3.13) holds then the nonlinear subspace iteration (3.3.8) and (3.3.9) converges (with the same choice of ω, spaces and mappings). This is summarised in the following Theorem:

Theorem 3.3.2 [[29], Theorem 1.7]

Assume **H1-H5** *and (3.3.13) hold and let* σ'_ω *be any value in the interval* $(\sigma_\omega, 1)$ *with*

σ_ω as given in (3.3.13). Then there is a neighbourhood U' of x^* such that the nonlinear subspace iteration (3.3.8) and (3.3.9) converges in U' with a convergence rate of σ'_ω.

This theorem is used to establish convergence of the ANF method for the drift-diffusion semiconductor equations as outlined in Section 3.2. To do this the key point is to establish (3.3.13). The proof relies heavily on graph theory which is discussed in Appendix A.

Remark 3.3.3 *Ortega and Rheinboldt [54, Section 10.3] also have results on the link between the convergence of linear iterations and corresponding nonlinear iterations.*

3.3.2 Convergence of the ANF Method Applied to the Semiconductor System for Small Applied Voltage

Some Preliminaries

It is proved in this section that the ANF method applied to the discretisation of the one dimensional drift-diffusion semiconductor equations converges for small applied voltage.

The differential equations to be solved are:

$$-\lambda^2\psi'' + \delta^2\{\exp(\psi - v) - \exp(w - \psi)\} - d = 0, \tag{3.3.14}$$

$$-(\exp(\psi - v)v')' - \sigma\rho_v r(\psi, v, w) = 0, \tag{3.3.15}$$

$$-(\exp(w - \psi)w')' + \sigma\rho_w r(\psi, v, w) = 0, \tag{3.3.16}$$

on the domain $\Omega = [0, 1]$. $\lambda, \delta, \sigma, \rho_v, \rho_w$ are positive constants, d is a (given) piecewise linear function and r is the generation/recombination term which we take to be:

$$r(\psi, v, w) = \frac{\exp(w - v) - 1}{\exp(w - \psi) + \exp(v - v) + 2}. \tag{3.3.17}$$

The boundary conditions on the system (3.3.14)-(3.3.16), with an applied voltage of V_0 at the left hand contact ($x = 0$ in the model) and V_1 at the right hand contact ($x = 1$), are:

$$\psi(0) = \sinh^{-1}\left(\frac{d(0)}{2\delta^2}\right) + \alpha_0, \qquad \psi(1) = \sinh^{-1}\left(\frac{d(1)}{2\delta^2}\right) + \alpha_1,$$

$$v(0) = \alpha_0, \qquad v(1) = \alpha_1,$$

$$w(0) = \alpha_0, \qquad w(1) = \alpha_1.$$

For the constant U_T given in Section 1.2, the $\alpha_i = V_i/U_T$, $i = 0, 1$, are the *scaled applied voltages.*

The finite element approximation to the solution of the system (3.3.14)-(3.3.16), with the given boundary conditions, is calculated on the grid $0 = x_0 < x_1 < \ldots < x_n < x_{n+1} = 1$, where n is the number of grid points in the interior of the domain. For each p we set $h_p := x_p - x_{p-1}$ and we define ϕ_p to be the standard hat function based on the mesh point p. The finite element method seeks approximate solutions Ψ, V and W:

$$\begin{aligned}
\Psi &= \left(\sinh^{-1}\left(\frac{d(0)}{2\delta^2}\right) + \alpha_0\right)\phi_0 + \sum_{p=1}^{n} \Psi_p \phi_p + \left(\sinh^{-1}\left(\frac{d(1)}{2\delta^2}\right) + \alpha_1\right)\phi_{n+1}, \\
V &= \alpha_0\phi_0 + \sum_{p=1}^{n} V_p \phi_p + \alpha_1 \phi_{n+1}, \\
W &= \alpha_0\phi_0 + \sum_{p=1}^{n} W_p \phi_p + \alpha_1 \phi_{n+1}
\end{aligned}$$

which are required to satisfy the equations:

$$\lambda^2(\Psi', \phi_p') + (\delta^2\{\exp(\Psi - V) - \exp(W - \Psi)\} - d, \phi_p) = 0, \qquad (3.3.18)$$

$$(\exp(\Psi - V)V', \phi_p') - \sigma\rho_v(r(\Psi, V, W), \phi_p) = 0, \qquad (3.3.19)$$

$$(\exp(W - \Psi)W', \phi_p') + \sigma\rho_w(r(\Psi, V, W), \phi_p) = 0, \qquad (3.3.20)$$

with p ranging over the interior points in the mesh.

The calculation of Ψ, V, W is equivalent to to the problem of finding the vector of unknowns: $\boldsymbol{X} := \left(\boldsymbol{\Psi}^T, \boldsymbol{V}^T, \boldsymbol{W}^T\right)^T$ in $\mathbb{R}^{3n}$, where $\boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W}$ contain the values of Ψ, V, W at the interior mesh points. It is also necessary to define $\tilde{\boldsymbol{\Psi}} \in \mathbb{R}^{n+2}$, this is the extended vector including the boundary values of Ψ, in the natural order. $\tilde{\boldsymbol{V}}, \tilde{\boldsymbol{W}}$ are defined analogously. Also define the vector of scaled voltages $\boldsymbol{\alpha} := (\alpha_0, \alpha_1)^T$.

To make the analysis simpler, the zero order terms in (3.3.18)-(3.3.20) are mass lumped (discussed in greater detail in Appendix B). This simply approximates a term (f, g) by its discrete counter part, $\langle f, g\rangle$, obtained using the trapezoidal rule:

$$\langle f, g\rangle := \sum_{p=1}^{n+1} h_p \left\{\frac{f(x_p)g(x_p) + f(x_{p-1})g(x_{p-1})}{2}\right\}. \qquad (3.3.21)$$

This means that the nonlinear zero order terms are approximated by diagonal nonlin-

earities in the discrete system, in an analogous fashion to the standard finite difference method.

After employing (3.3.21), the finite element system (3.3.18)-(3.3.20) can be written in the form:

$$\mathbf{F}\left(\left(\boldsymbol{\Psi}^T, \boldsymbol{V}^T, \boldsymbol{W}^T\right)^T, \boldsymbol{\alpha}\right) = \begin{bmatrix} \lambda^2 \tilde{A}(0)\tilde{\boldsymbol{\Psi}} + \boldsymbol{e}(\boldsymbol{\Psi} - \boldsymbol{V}) - \boldsymbol{e}(\boldsymbol{W} - \boldsymbol{\Psi}) - \boldsymbol{d} \\ \tilde{A}(\tilde{\boldsymbol{\Psi}} - \tilde{\boldsymbol{V}})\tilde{\boldsymbol{V}} - \rho_v \boldsymbol{r}(\boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W}) \\ \tilde{A}(\tilde{\boldsymbol{W}} - \tilde{\boldsymbol{\Psi}})\tilde{\boldsymbol{W}} + \rho_w \boldsymbol{r}(\boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W}) \end{bmatrix} = \mathbf{0}. \tag{3.3.22}$$

The matrices in (3.3.22) are defined by:

$$\begin{aligned} \tilde{A}(\boldsymbol{B})_{pq} &= \left(\exp\left(\sum_{r=0}^{n+1} B_r \phi_r\right) \nabla\phi_p, \nabla\phi_q\right), \quad p = 1, 2, \ldots, n, \ q = 0, 1, \ldots, n+1, \\ &\qquad\qquad \boldsymbol{B} \in \mathbb{R}^{n+2}, \\ e(\boldsymbol{B})_p &= \delta^2 \left(\frac{h_p + h_{p+1}}{2}\right) \exp(B_p), \quad p = 1, 2, \ldots, n, \ \boldsymbol{B} \in \mathbb{R}^n, \\ d_p &= \left(\frac{h_p + h_{p+1}}{2}\right) d(x_p), \quad p = 1, 2, \ldots, n, \\ r(\boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W})_p &= \sigma \left(\frac{h_p + h_{p+1}}{2}\right) r(\Psi_p, V_p, W_p), \quad p = 1, 2, \ldots, n. \end{aligned}$$

For a given vector of scaled voltages $\boldsymbol{\alpha}$, let $\boldsymbol{X}(\boldsymbol{\alpha}) = \left(\boldsymbol{\Psi}_\alpha^T, \boldsymbol{V}_\alpha^T, \boldsymbol{W}_\alpha^T\right)^T$ denote the solution to $\mathbf{F}\left(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha}\right) = \mathbf{0}$, where the nonlinear finite element system $\mathbf{F}$ is given by (3.3.22). We showed in Chapter 2 that for $\boldsymbol{\alpha}$ sufficiently small there exists a unique solution $\boldsymbol{X}(\boldsymbol{\alpha})$ to (3.3.22).

Before considering the convergence of the ANF method applied to the semiconductor system we first introduce a standard nonlinear solver for device modelling which we will compare with the ANF procedure at the end of this chapter.

3.3.3 Gummel's Method

"Gummel's Method" is the name given in the semiconductor literature for a group of nonlinear block Gauss-Seidel algorithms solving the discrete drift diffusion equations. The method was first introduced by Gummel in 1964 ([36]) and is still extensively used in modern semiconductor device modelling - see for example [8], [9] and [55]. The method has been extensively analysised in the context of the undiscretised equations, see for example [43], [44], and in the discrete case in [23].

The variant of the algorithm which we use (written in the original, rather than the discretised, variables for clarity) is:

1. Make initial guesses at the solutions to the semiconductor equations: ψ^0, v^0 and w^0.

2. For $k = 0, 1, \ldots$, iterate the following three steps

   (a) Solve using Newton's method:

$$-\lambda^2 \Delta \psi^{k+1} + \delta^2 \{\exp(\psi^{k+1} - v^k) - \exp(w^k - \psi^{k+1})\} - d = 0, \quad \text{for } \psi^{k+1}.$$

   (b) Solve the linear system:

$$-\nabla.(\exp(\psi^{k+1} - v^k)\nabla v^{k+1}) - \sigma\rho_v r(\psi^{k+1}, v^k, w^k) = 0, \quad \text{for } v^{k+1}.$$

   (c) Solve the linear system:

$$-\nabla.(\exp(w^k - \psi^{k+1})\nabla w^{k+1}) + \sigma\rho_w r(\psi^{k+1}, v^{k+1}, w^k) = 0, \quad \text{for } w^{k+1}.$$

Remark 3.3.4 *An alternative Gummel method would be to use Newton's method to solve the semilinear system:*

$$-\nabla.(\exp(\psi^{k+1} - v^k)\nabla v^{k+1}) - \sigma\rho_v r(\psi^{k+1}, v^{k+1}, w^k) = 0,$$

for v^{k+1} in step 2b and the analogous semilinear system in step 2c. This method is also known as a Gauss-Seidel Newton method in the literature.

Numerical Results

In this section the Gummel's method described above is applied to the finite element discretisation of the system modelling a PN diode in one dimension. Results show that the method applied to this semiconductor system converges for sufficiently small applied voltage. It turns out that Gummel's method, with the same initial guess strategy, is more robust to high voltages than Newton's method.

The finite element method with mass lumping is applied to the semiconductor system (3.3.14)-(3.3.16) in one dimension with $\Omega = [0, 1]$. Since the aim is to model a PN

Number of interior mesh points	Voltage applied at 0	Voltage applied at 1	Number of Gummel iterations
9	0	0.1	11
9	0	0.3	13
9	0	0.4	Diverges
9	5.4	5.5	11
9	0.1	0	11
9	0.5	0	12
9	0.9	0	16
9	1.0	0	Diverges
21	0	0.1	17
21	0	0.4	Diverges
21	0.1	0	17
21	0.9	0	24
21	1.0	0	Diverges

Table 3.1: Gummel's method applied to the one dimensional discretisation of a PN diode on a uniform grid. The iteration is stopped when the change in $\left(\boldsymbol{\Psi}^T, \boldsymbol{V}^T, \boldsymbol{W}^T\right)^T$ is less than 5×10^{-5}.

diode the doping profile chosen is -1 on $[0, 1/2)$ and +1 on $(1/2, 1]$. The equations are discretised with respect to a uniform grid with n interior mesh points. The initial guess for ψ, v and w is the doping profile scaled to match the relevant Dirichlet boundary conditions.

The results for Gummel's method are contained in Table 3.1. The method only converges for small reverse bias (the voltage applied to the contact at the n-type region is more positive than the voltage applied to the contact at the p-type region, as discussed in Chapter 1). but converges for much larger forward bias. We will see that this behaviour is a special feature of Gummel's method.

Gummel's method is a decoupled method and neglects the coupling between the PDEs. As discussed in the introduction to this chapter it has been suggested in [8] that, for large applied voltages, the coupling between the PDEs is much stronger than the coupling between the values of the solution of a single PDE at different points of the mesh. We believe this is the reason Gummel's method breaks down so quickly in reverse bias. The ANF method considered in this chapter, which takes into account the coupling between the PDEs at each iterative step, will be shown to converge for a greater range of applied voltages for the same test problem.

3.3.4 Convergence of the ANF Method Continued

In order to prove that the ANF method converges for small applied voltage we shall show it converges for a zero applied voltage ($\boldsymbol{\alpha} = \mathbf{0}$) and then use a perturbation argument to extend the proof to small applied voltages. For the zero voltage proof it is necessary to consider in detail the solution to (3.3.22) when $\boldsymbol{\alpha} = \mathbf{0}$. This is easily shown to be the vector:

$$\boldsymbol{X}(\mathbf{0}) = \left[\boldsymbol{\Psi}_0^T, \mathbf{0}^T, \mathbf{0}^T\right]^T \in \mathbb{R}^{3n}, \tag{3.3.23}$$

where $\boldsymbol{\Psi}_0$ consists of the nodal values of the finite element solution:

$$\Psi_0 = \left(\sinh^{-1}\left(\frac{d(0)}{2\delta^2}\right)\right)\phi_0 \; + \; \sum_{p=1}^{n} (\Psi_0)_p \, \phi_p \; + \; \left(\sinh^{-1}\left(\frac{d(1)}{2\delta^2}\right)\right)\phi_{n+1},$$

of the system

$$\lambda^2(\Psi_0', \phi_p') + \langle 2\delta^2 \sinh(\Psi_0) - d, \phi_p \rangle = 0, \quad p = 1, 2, \ldots, n. \tag{3.3.24}$$

The Jacobian of (3.3.22) with respect to $\boldsymbol{X}$, at $(\boldsymbol{X}(\mathbf{0}), \mathbf{0})$, is given by:

$$\mathbf{F}_X\left(\boldsymbol{X}(\mathbf{0}), \mathbf{0}\right) = \begin{bmatrix} \lambda^2 A(0) + E(\boldsymbol{\Psi}_0) + E(-\boldsymbol{\Psi}_0) & -E(\boldsymbol{\Psi}_0) & -E(-\boldsymbol{\Psi}_0) \\ 0 & A(\tilde{\boldsymbol{\Psi}}_0) + \rho_v H(\boldsymbol{\Psi}_0) & -\rho_v H(\boldsymbol{\Psi}_0) \\ 0 & -\rho_w H(\boldsymbol{\Psi}_0) & A(-\tilde{\boldsymbol{\Psi}}_0) + \rho_w H(\boldsymbol{\Psi}_0) \end{bmatrix}, \tag{3.3.25}$$

where $A(\boldsymbol{B})$ consists of the matrix $\tilde{A}(\boldsymbol{B})$ minus its first and last columns and, defining $\boldsymbol{\delta}_{pq}$ to be the Kronecker delta, the matrices E and H are defined to be:

$$\begin{aligned} \mathrm{E}(\boldsymbol{B})_{pq} &= \delta^2 \left(\frac{h_p + h_{p+1}}{2}\right) \exp(B_p)\boldsymbol{\delta}_{pq}, \quad p, q = 1, 2, \ldots, n, \quad \boldsymbol{B} \in \mathbb{R}^n \\ \mathrm{H}(\boldsymbol{B})_{pq} &= \left(\frac{h_p + h_{p+1}}{2}\right) \frac{\sigma}{[2\cosh(B_p) + 2]} \boldsymbol{\delta}_{pq}, \quad p, q = 1, 2, \ldots, n, \quad \boldsymbol{B} \in \mathbb{R}^n. \end{aligned}$$

Remark 3.3.5 *Since the Jacobian matrix, (3.3.25), will be referred to frequently, it is useful to have shorthand notation for the entries of the matrices A, E and H. The*

matrix $A(\boldsymbol{B})$ *is a tridiagonal matrix which has entries:*

$$A(\boldsymbol{B})_{pq} = \begin{cases} -a_p(\boldsymbol{B}), & q = p-1 \\ a_p(\boldsymbol{B}) + a_{p+1}(\boldsymbol{B}), & q = p \\ -a_{p+1}(\boldsymbol{B}), & q = p+1 \\ 0, & otherwise \end{cases},$$

where,

$$a_p(\boldsymbol{B}) = \frac{1}{h_p}\left(\frac{\exp(B_p) - \exp(B_{p-1})}{B_p - B_{p-1}}\right), \quad if\ B_{p-1} \neq B_p$$

and

$$a_p(\boldsymbol{B}) = \frac{1}{h_p}\exp(B_p), \quad if\ B_{p-1} = B_p.$$

$E(\boldsymbol{B})$ *is a diagonal matrix which has entries:*

$$E(\boldsymbol{B})_{pq} = \begin{cases} e_p(\boldsymbol{B}), & q = p \\ 0, & otherwise \end{cases},$$

where

$$e_p(\boldsymbol{B}) = \delta^2\left(\frac{h_p + h_{p+1}}{2}\right)\exp(B_p).$$

Similarly $H(\boldsymbol{B})$ *is a diagonal matrix with entries:*

$$H(\boldsymbol{B})_{pq} = \begin{cases} h_p(\boldsymbol{B}), & q = p \\ 0, & otherwise \end{cases},$$

where

$$h_p(\boldsymbol{B}) = \left(\frac{h_p + h_{p+1}}{2}\right)\frac{\sigma}{[2\cosh(B_p) + 2]}.$$

Before considering the convergence of the ANF method described in Section 3.2, it is first necessary to put it in the framework of the nonlinear subspace iteration method of Section 3.3.1. To do this we must define the spaces X, X_κ and the linear injective and surjective mappings, p_κ and r_κ. The spaces are:

$$X = \mathbb{R}^{3n} \quad \text{and} \quad X_\kappa = \mathbb{R}^3, \quad \kappa = 1, 2, \ldots, n.$$

Since the ANF method is simply a block Jacobi method with the matrix blocked by

mesh point, the linear injective mappings $p_\kappa : X_\kappa \to X$ are defined by:

$$\left(p_\kappa \begin{bmatrix} a \\ b \\ c \end{bmatrix} \right)_l = \begin{cases} a, & l = \kappa \\ b, & l = \kappa + n \\ c, & l = \kappa + 2n \\ 0, & \text{otherwise} \end{cases}, \quad \kappa = 1, 2, \ldots, n, \quad l = 1, 2, \ldots, 3n. \tag{3.3.26}$$

That is p_κ extends a vector $(a, b, c)^T \in \mathbb{R}^3$ to a vector in $\mathbb{R}^{3n}$ by placing a in the κth entry of the image, b in the $\kappa + n$th entry and c in the $\kappa + 2n$th entry and taking zeros elsewhere.

The linear surjective mappings $r_\kappa : X \to X_\kappa$ are taken to be the transposes of p_κ, which means that the r_κ's are given by:

$$r_\kappa \boldsymbol{Y} = \begin{bmatrix} Y_\kappa \\ Y_{\kappa+n} \\ Y_{\kappa+2n} \end{bmatrix}, \quad \kappa = 1, 2, \ldots, n, \tag{3.3.27}$$

here Y_κ is the κth entry of the vector $\boldsymbol{Y} \in X$. With these definitions it is clear that $X = \sum_{\kappa=1}^{n} p_\kappa X_\kappa$, as required by the Dryja-Hackbusch theory of Section 3.3.1.

The proof of convergence

In this part the following theorem is proved:

Theorem 3.3.6 *There exists an $r > 0$ such that for all scaled applied voltages $\boldsymbol{\alpha} \in \mathcal{B}(\mathbf{0}, r)$, the ANF iteration applied to the nonlinear system (3.3.22) converges.*

Theorem 3.3.6 is proved in three stages.

Stage I First it is shown that the ANF iteration applied to the linearised system converges with zero applied voltage.

Stage II A perturbation argument (with respect to voltage) is used to prove that the ANF iteration applied to the linearised system converges for small voltage.

Stage III Finally an application of Theorem 3.3.2 shows that the ANF method applied to the nonlinear system (3.3.22), with small applied voltage, converges.

Stage I

To obtain the proof of this part recall the linear subspace iteration (3.3.11) and observe that, with p_κ and r_κ defined in (3.3.26) and the discussion following it, (3.3.11) is equivalent to the damped block Jacobi method applied to the system:

$$\tilde{A}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}} \tag{3.3.28}$$

where the relation between the matrices A and $\tilde{A}$ and vectors $\boldsymbol{x}, \tilde{\boldsymbol{x}}, \boldsymbol{b}$ and $\tilde{\boldsymbol{b}}$ is described in Section 3.2.1. Thus the ANF method corresponds to $\omega = 1$ (an undamped subspace iteration) and may be written more simply as

$$\tilde{\boldsymbol{x}}^{l+1} = (I - \tilde{D}^{-1}\tilde{A})\tilde{\boldsymbol{x}}^{l} + \tilde{D}^{-1}\tilde{\boldsymbol{b}} \tag{3.3.29}$$

where $\tilde{D}$ denotes the block diagonal of $\tilde{A}$. It is well known (see for example [54]) that the iteration (3.3.29) converges provided that the iteration matrix:

$$M_0 := (I - \tilde{D}^{-1}\tilde{A}) \tag{3.3.30}$$

has eigenvalues which are less than one in modulus.

Thus to obtain the first stage of Theorem 3.3.6 we must prove that M_0 has spectral radius less than one, when A is the Jacobian of the discretised semiconductor system (3.3.22) at the zero applied voltage solution $(\boldsymbol{X}(\mathbf{0}), \mathbf{0})$. For the rest of this stage of the proof M_0 is given by (3.3.30) with A given by (3.3.25) and $\tilde{A}$ the re-blocked matrix (3.2.6). A method of proof similar to those given in [37, Theorem 6.4.10] and [69, Theorem 3.4] is used. These theorems are concerned with the convergence of the Jacobi iteration, here we apply the ideas to the block version.

The proof uses a matrix $|M_0|$, the 'modulus' matrix of M_0, defined by:

$$(|M_0|)_{pq} = |\,(M_0)_{pq}\,|, \quad p, q = 1, 2, \ldots, n.$$

Recalling that $B \leq A$ means that all the entries of B are less than or equal in size to the corresponding entries in A, it is known that $\rho(B) \leq \rho(A)$ whenever $|B| \leq A$ (see for example [37, Exercise 6.3.11] or [74, Theorem 1.16, Chapter 2]). So, if it can be shown that the spectral radius of $|M_0|$ is less that one, it follows that the spectral radius of

M_0 is less than one and that (3.3.29) converges.

In order to check that the spectral radius of $|M_0|$ is less than one it is necessary to look more closely at the structure of $\tilde{A}$ and $|M_0|$. First recall that $\tilde{D}$ is block diagonal with diagonal blocks equal to the diagonal blocks of $\tilde{A}$ in (3.2.6). From (3.3.25) these are easily calculated to be

$$\tilde{D}_\kappa = \tilde{A}_{\kappa,\kappa} = \begin{bmatrix} \frac{\lambda^2}{h_\kappa} + \frac{\lambda^2}{h_{\kappa+1}} + & -e_\kappa(\boldsymbol{\Psi}_0) & -e_\kappa(-\boldsymbol{\Psi}_0) \\ e_\kappa(\boldsymbol{\Psi}_0) + e_\kappa(-\boldsymbol{\Psi}_0) & & \\ 0 & a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0) + & -\rho_v h_\kappa(\boldsymbol{\Psi}_0) \\ & \rho_v h_\kappa(\boldsymbol{\Psi}_0) & \\ 0 & -\rho_w h_\kappa(\boldsymbol{\Psi}_0) & a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0) + \\ & & \rho_w h_\kappa(\boldsymbol{\Psi}_0) \end{bmatrix}. \tag{3.3.31}$$

If we introduce the diagonal matrix

$$O_\kappa = \begin{bmatrix} \frac{-\lambda^2}{h_\kappa} & 0 & 0 \\ 0 & -a_\kappa(\tilde{\boldsymbol{\Psi}}_0) & 0 \\ 0 & 0 & -a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) \end{bmatrix}, \tag{3.3.32}$$

then:

$$\tilde{A} = \begin{bmatrix} \tilde{D}_1 & O_2 & & & \\ O_2 & \tilde{D}_2 & O_3 & & \\ & O_3 & \ddots & \ddots & \\ & & \ddots & \tilde{D}_{n-1} & O_n \\ & & & O_n & \tilde{D}_n \end{bmatrix} \tag{3.3.33}$$

and $\tilde{D}$ is the block diagonal of $\tilde{A}$, so that:

$$M_0 = \begin{bmatrix} 0 & \left(\tilde{D}_1\right)^{-1} O_2 & & & \\ \left(\tilde{D}_2\right)^{-1} O_2 & 0 & \left(\tilde{D}_2\right)^{-1} O_3 & & \\ & \left(\tilde{D}_3\right)^{-1} O_3 & \ddots & \ddots & \\ & & \ddots & 0 & \left(\tilde{D}_{n-1}\right)^{-1} O_n \\ & & & \left(\tilde{D}_n\right)^{-1} O_n & 0 \end{bmatrix}. \tag{3.3.34}$$

In the above 0 represents the 3×3 zero matrix.

To show that the spectral radius of $|M_0|$ is less than one, we examine its row sums:

Lemma 3.3.7 *The first three and last three rows of $|M_0|$ have row sum strictly less than one, all other rows of $|M_0|$ have row sum equal to one.*

Proof

First it is shown that the three row sums of the modulus matrix:

$$|\left(\tilde{D}_\kappa\right)^{-1} O_\kappa| + |\left(\tilde{D}_\kappa\right)^{-1} O_{\kappa+1}| \tag{3.3.35}$$

are all equal to one for $\kappa = 2, 3, \ldots, n-1$. These row sums correspond to the row sums of all the rows of $|M_0|$, except the first three and last three.

For the purpose of the proof the following short-hand notation is introduced:

$$\tilde{D}_\kappa = \begin{bmatrix} D1_\kappa & D2_\kappa & D3_\kappa \\ 0 & D4_\kappa & D5_\kappa \\ 0 & D6_\kappa & D7_\kappa \end{bmatrix}$$

and

$$O_\kappa = \begin{bmatrix} O1_\kappa & 0 & 0 \\ 0 & O2_\kappa & 0 \\ 0 & 0 & O3_\kappa \end{bmatrix}.$$

Comparing this new notation with (3.3.31) and (3.3.32) it can be seen that, $D1_\kappa = \frac{\lambda^2}{h_\kappa} + \frac{\lambda^2}{h_{\kappa+1}} + e_\kappa(\Psi_0) + e_\kappa(-\Psi_0)$, etc.

With this notation

$$\left(\tilde{D}_\kappa\right)^{-1} O_\kappa = \begin{bmatrix} \frac{O1_\kappa}{D1_\kappa} & \frac{O2_\kappa(D7_\kappa D2_\kappa - D6_\kappa D3_\kappa)}{D1_\kappa(D5_\kappa D6_\kappa - D7_\kappa D4_\kappa)} & \frac{O3_\kappa(D4_\kappa D3_\kappa - D5_\kappa D2_\kappa)}{D1_\kappa(D5_\kappa D6_\kappa - D7_\kappa D4_\kappa)} \\ 0 & \frac{-O2_\kappa D7_\kappa}{(D5_\kappa D6_\kappa - D7_\kappa D4_\kappa)} & \frac{O3_\kappa D5_\kappa}{(D5_\kappa D6_\kappa - D7_\kappa D4_\kappa)} \\ 0 & \frac{O2_\kappa D6_\kappa}{(D5_\kappa D6_\kappa - D7_\kappa D4_\kappa)} & \frac{-O3_\kappa D4_\kappa}{(D5_\kappa D6_\kappa - D7_\kappa D4_\kappa)} \end{bmatrix}. \tag{3.3.36}$$

Note that in (3.3.38) below we shall show that $D5_\kappa D6_\kappa - D7_\kappa D4_\kappa \neq 0$.

Finding the row sums of the modulus matrix associated with (3.3.35) is equivalent

to finding the row sums of the matrix:

$$\begin{bmatrix} \frac{|O1_\kappa|+|O1_{\kappa+1}|}{|D1_\kappa|} & \frac{(|O2_\kappa|+|O2_{\kappa+1}|)|D7_\kappa D2_\kappa - D6_\kappa D3_\kappa|}{|D1_\kappa||D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|} & \frac{(|O3_\kappa|+|O3_{\kappa+1}|)|D4_\kappa D3_\kappa - D5_\kappa D2_\kappa|}{|D1_\kappa||D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|} \\ 0 & \frac{(|O2_\kappa|+|O2_{\kappa+1}|)|D7_\kappa|}{|D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|} & \frac{(|O3_\kappa|+|O3_{\kappa+1}|)|D5_\kappa|}{|D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|} \\ 0 & \frac{(|O2_\kappa|+|O2_{\kappa+1}|)|D6_\kappa|}{|D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|} & \frac{(|O3_\kappa|+|O3_{\kappa+1}|)|D4_\kappa|}{|D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|} \end{bmatrix}. \tag{3.3.37}$$

First consider the sum of the third row of the matrix (3.3.37):

$$\frac{(|O2_\kappa| + |O2_{\kappa+1}|)|D6_\kappa| + (|O3_\kappa| + |O3_{\kappa+1}|)|D4_\kappa|}{|D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|}.$$

Returning once again to the notation of (3.3.31) and (3.3.32) it is clear that:

$$\begin{aligned} &(|O2_\kappa| + |O2_{\kappa+1}|)|D6_\kappa| + (|O3_\kappa| + |O3_{\kappa+1}|)|D4_\kappa| \\ &\quad = \rho_w \left(a_\kappa(\tilde{\Psi}_0) + a_{\kappa+1}(\tilde{\Psi}_0)\right) h_\kappa(\Psi_0) \; + \\ &\qquad \left(a_\kappa(\tilde{\Psi}_0) + a_{\kappa+1}(\tilde{\Psi}_0) + \rho_v h_\kappa(\Psi_0)\right) \left(a_\kappa(-\tilde{\Psi}_0) + a_{\kappa+1}(-\tilde{\Psi}_0)\right) \\ &\quad = h_\kappa(\Psi_0) \left\{\rho_v \left(a_\kappa(-\tilde{\Psi}_0) + a_{\kappa+1}(-\tilde{\Psi}_0)\right) + \rho_w \left(a_\kappa(\tilde{\Psi}_0) + a_{\kappa+1}(\tilde{\Psi}_0)\right)\right\} + \\ &\qquad \left\{a_\kappa(\tilde{\Psi}_0) + a_{\kappa+1}(\tilde{\Psi}_0)\right\} \left\{a_\kappa(-\tilde{\Psi}_0) + a_{\kappa+1}(-\tilde{\Psi}_0)\right\} \end{aligned}$$

and that

$$\begin{aligned} &|D5_\kappa D6_\kappa - D7_\kappa D4_\kappa| \\ &\quad = |-\left[a_\kappa(\tilde{\Psi}_0) + a_{\kappa+1}(\tilde{\Psi}_0) + \rho_v h_\kappa(\Psi_0)\right] \times \\ &\qquad \left[a_\kappa(-\tilde{\Psi}_0) + a_{\kappa+1}(-\tilde{\Psi}_0) + \rho_w h_\kappa(\Psi_0)\right] \; + \; \rho_v \rho_w \left(h_\kappa(\Psi_0)\right)^2 | \\ &\quad = |-\left[a_\kappa(\tilde{\Psi}_0) + a_{\kappa+1}(\tilde{\Psi}_0)\right] \left[a_\kappa(-\tilde{\Psi}_0) + a_{\kappa+1}(-\tilde{\Psi}_0)\right] - \\ &\qquad \rho_w h_\kappa(\Psi_0) \left[a_\kappa(\tilde{\Psi}_0) + a_{\kappa+1}(\tilde{\Psi}_0)\right] - \\ &\qquad \rho_v h_\kappa(\Psi_0) \left[a_\kappa(-\tilde{\Psi}_0) + a_{\kappa+1}(-\tilde{\Psi}_0)\right] | \\ &\quad = h_\kappa(\Psi_0) \left\{\rho_v \left(a_\kappa(-\tilde{\Psi}_0) + a_{\kappa+1}(-\tilde{\Psi}_0)\right) + \rho_w \left(a_\kappa(\tilde{\Psi}_0) + a_{\kappa+1}(\tilde{\Psi}_0)\right)\right\} + \\ &\qquad \left\{a_\kappa(\tilde{\Psi}_0) + a_{\kappa+1}(\tilde{\Psi}_0)\right\} \left\{a_\kappa(-\tilde{\Psi}_0) + a_{\kappa+1}(-\tilde{\Psi}_0)\right\}. \end{aligned} \tag{3.3.38}$$

Therefore

$$\frac{(|O2_\kappa| + |O2_{\kappa+1}|)|D6_\kappa| + (|O3_\kappa| + |O3_{\kappa+1}|)|D4_\kappa|}{|D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|} = 1$$

and the row sum of the third row of (3.3.37) is equal to one.

Note: From (3.3.38) it can be seen that $|D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|$ is never equal to zero since $a_\kappa(\boldsymbol{B})$ and $h_\kappa(\boldsymbol{B})$ are always positive for all vectors $\boldsymbol{B}$ (see Remark 3.3.5).

Next consider the second row sum of (3.3.37):

$$\frac{(|O2_\kappa| + |O2_{\kappa+1}|)|D7_\kappa| + (|O3_\kappa| + |O3_{\kappa+1}|)|D5_\kappa|}{|D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|}.$$

Again returning to the notation used in (3.3.31) and (3.3.32) it can be seen that

$$\begin{aligned}
&(|O2_\kappa| + |O2_{\kappa+1}|)|D7_\kappa| + (|O3_\kappa| + |O3_{\kappa+1}|)|D5_\kappa| \\
&\quad = \left[a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0)\right]\left[a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0) + \rho_w h_\kappa(\boldsymbol{\Psi}_0)\right] + \\
&\qquad \left[a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0)\right]\rho_v h_\kappa(\boldsymbol{\Psi}_0) \\
&\quad = \left\{a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0)\right\}\left\{a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0)\right\} + \\
&\qquad h_\kappa(\boldsymbol{\Psi}_0)\left\{\rho_v\left(a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0)\right) + \rho_w\left(a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0)\right)\right\} \\
&\quad = |D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|.
\end{aligned}$$

The last line follows from (3.3.38).

Thus the sum of the second row of (3.3.37) is also equal to one.

Finally considering the first row of (3.3.37) it is required to calculate the size of

$$\frac{|O1_\kappa| + |O1_{\kappa+1}|}{|D1_\kappa|} + \frac{(|O2_\kappa| + |O2_{\kappa+1}|)|D7_\kappa D2_\kappa - D6_\kappa D3_\kappa| + (|O3_\kappa| + |O3_{\kappa+1}|)|D4_\kappa D3_\kappa - D5_\kappa D2_\kappa|}{|D1_\kappa||D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|}.$$

First consider:

$$\begin{aligned}
&(|O2_\kappa| + |O2_{\kappa+1}|)|D7_\kappa D2_\kappa - D6_\kappa D3_\kappa| + (|O3_\kappa| + |O3_{\kappa+1}|)|D4_\kappa D3_\kappa - D5_\kappa D2_\kappa| \\
&\quad = \left\{\left|\left[-e_\kappa(\boldsymbol{\Psi}_0)\left\{a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0) + \rho_w h_\kappa(\boldsymbol{\Psi}_0)\right\} - \rho_w e_\kappa(-\boldsymbol{\Psi}_0)h_\kappa(\boldsymbol{\Psi}_0)\right]\right| \times \right. \\
&\qquad \left.\left[a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0)\right]\right\} + \left\{\left[a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0)\right] \times \right. \\
&\qquad \left.\left|\left[-e_\kappa(-\boldsymbol{\Psi}_0)\left\{a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0) + \rho_v h_\kappa(\boldsymbol{\Psi}_0)\right\} - \rho_v e_\kappa(\boldsymbol{\Psi}_0)h_\kappa(\boldsymbol{\Psi}_0)\right]\right|\right\} \\
&\quad = e_\kappa(\boldsymbol{\Psi}_0)\left[a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0)\right]\left[a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0)\right] + \\
&\qquad \rho_w h_\kappa(\boldsymbol{\Psi}_0)\left[e_\kappa(\boldsymbol{\Psi}_0) + e_\kappa(-\boldsymbol{\Psi}_0)\right]\left[a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0)\right] + \\
&\qquad e_\kappa(-\boldsymbol{\Psi}_0)\left[a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0)\right]\left[a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0)\right] +
\end{aligned}$$

$$
\begin{aligned}
& \rho_v h_\kappa(\Psi_0)\left[e_\kappa(\Psi_0)+e_\kappa(-\Psi_0)\right]\left[a_\kappa(-\tilde{\Psi}_0)+a_{\kappa+1}(-\tilde{\Psi}_0)\right] \\
= \; & \left[e_\kappa(\Psi_0)+e_\kappa(-\Psi_0)\right]\left\{\left[a_\kappa(\tilde{\Psi}_0)+a_{\kappa+1}(\tilde{\Psi}_0)\right]\left[a_\kappa(-\tilde{\Psi}_0)+a_{\kappa+1}(-\tilde{\Psi}_0)\right]+\right. \\
& \left. h_\kappa(\Psi_0)\left[\rho_v\left(a_\kappa(-\tilde{\Psi}_0)+a_{\kappa+1}(-\tilde{\Psi}_0)\right)+\rho_w\left(a_\kappa(\tilde{\Psi}_0)+a_{\kappa+1}(\tilde{\Psi}_0)\right)\right]\right\} \\
= \; & \left[e_\kappa(\Psi_0)+e_\kappa(-\Psi_0)\right]|D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|.
\end{aligned}
$$

From the above it follows that

$$
\begin{aligned}
& \frac{(|O2_\kappa|+|O2_{\kappa+1}|)|D7_\kappa D2_\kappa - D6_\kappa D3_\kappa| + (|O3_\kappa|+|O3_{\kappa+1}|)|D4_\kappa D3_\kappa - D5_\kappa D2_\kappa|}{|D1_\kappa||D5_\kappa D6_\kappa - D7_\kappa D4_\kappa|} \\
& \qquad + \frac{|O1_\kappa|+|O1_{\kappa+1}|}{|D1_\kappa|} \\
= \; & \frac{1}{|D1_\kappa|}\left[|O1_\kappa|+|O1_{\kappa+1}|+\left[e_\kappa(\Psi_0)+e_\kappa(-\Psi_0)\right]\right] \\
= \; & \frac{1}{|D1_\kappa|}\left[\frac{\lambda^2}{h_\kappa}+\frac{\lambda^2}{h_{\kappa+1}}+\left[e_\kappa(\Psi_0)+e_\kappa(-\Psi_0)\right]\right] \\
= \; & 1
\end{aligned}
$$

by the definition of $D1_\kappa$.

Thus the row sums of the matrix (3.3.37) are all equal to one. Separating (3.3.37) into the parts corresponding to $\left(\tilde{D}_\kappa\right)^{-1} O_\kappa$ and $\left(\tilde{D}_\kappa\right)^{-1} O_{\kappa+1}$ and padding with zeros in an appropriate way gives the result that the maximum row sum of the matrix $|M_0|$ is one. However the first and last three rows of $|M_0|$ contain only one of the modulus matrices corresponding to $\left(\tilde{D}_\kappa\right)^{-1} O_\kappa$ and $\left(\tilde{D}_\kappa\right)^{-1} O_{\kappa+1}$ and since (where addition and equality are understood in the sense of row sums)

$$
\left|\left(\tilde{D}_\kappa\right)^{-1} O_\kappa\right| + \left|\left(\tilde{D}_\kappa\right)^{-1} O_{\kappa+1}\right| = 1, \quad \text{for all } \kappa
$$

it follows that the row sum of the first and last three rows will be less than one. Proving the lemma. □

This lemma is used to prove that the spectral radii of $|M_0|$ and M_0 are less than one. the result is contained in the next Lemma:

Lemma 3.3.8

i *The $3n \times 3n$ matrix $|M_0|$ has a spectral radius less than one.*

ii *The linear ANF iteration matrix, M_0. also has spectral radius less than one.*

iii *The linear ANF method applied to the linearisation of (3.3.22) at the zero current solution converges.*

Proof The graph theory terms used in this proof are defined in Appendix A.

By construction $|M_0| \geq 0$. Hence there exists (by the Perron-Frobenius Theorem [66, Chapter 5]) an eigenvector $\boldsymbol{x} \geq 0$ with $\|\boldsymbol{x}\|_\infty = 1$ such that the eigenvalue, λ, associated with $\boldsymbol{x}$, is equal to the spectral radius of $|M_0|$. To prove (i) we must show $\lambda < 1$. Let $\alpha \in \{1, 2, \ldots, 3n\}$ be an index such that $x_\alpha = 1$.

First we claim that

$$\text{either } \lambda < 1 \text{ or } x_\gamma = 1 \text{ for all } \gamma \in \mathcal{G}_\alpha. \tag{3.3.39}$$

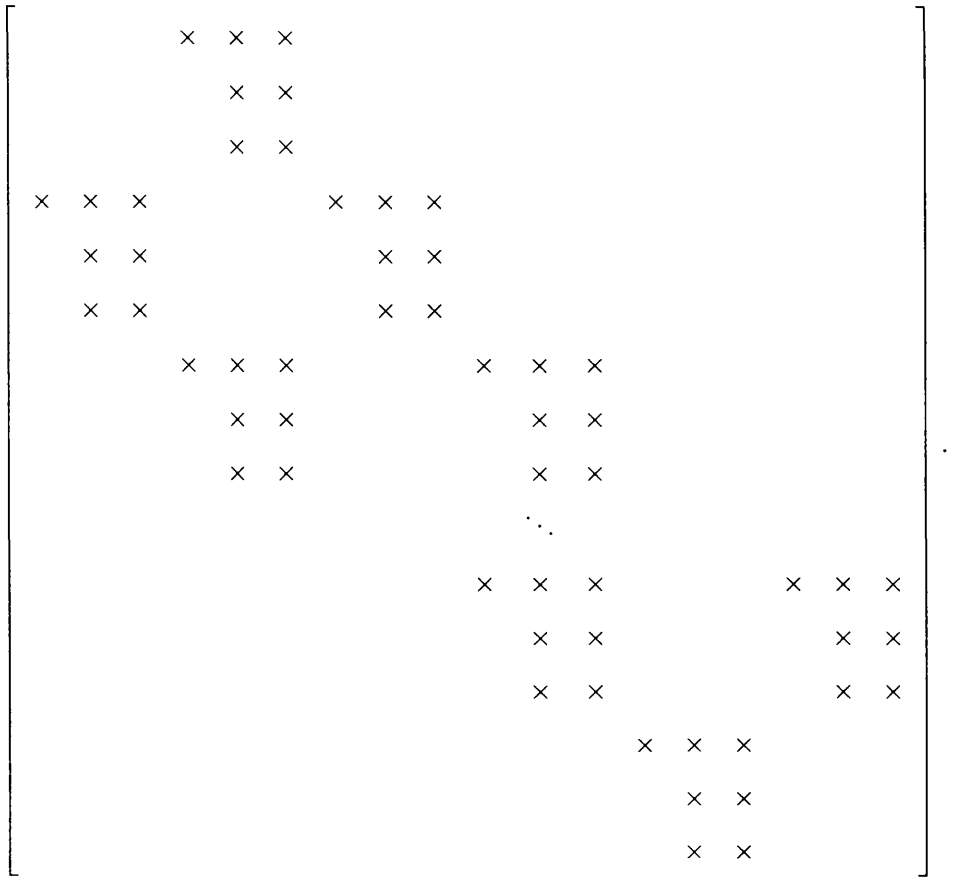
This claim follows by induction provided we show that either $\lambda < 1$ or $x_\gamma = 1$ for all γ, such that α is *directly* connected to γ. To show this, denote the elements of $|M_0|$ by $|m|_{ij}$, i.e. $|M_0| = (|m|_{ij})$. Then, since it is known that $\sum_{j=1}^{3n} |m|_{ij} \leq 1$ for all $i = 1, 2, \ldots, 3n$ (from Lemma 3.3.7), it follows that since $\boldsymbol{x}$ is an eigenvalue of $|M_0|$ with eigenvalue λ:

$$\lambda = \lambda x_\alpha = (|M_0|\boldsymbol{x})_\alpha = \sum_{j=1}^{3n} |m|_{\alpha j} x_j \leq \sum_{j=1}^{3n} |m|_{\alpha j} \leq 1.$$

Since λ can only be equal to one if $x_\gamma = 1$ for all γ with $|m|_{\alpha\gamma} \neq 0$ (i.e. all γ such that α is directly connected to γ), the claim (3.3.39) follows by induction.

To finish the proof of part (i) we show that $\lambda < 1$ even if $x_\gamma = 1$ for all $\gamma \in \mathcal{G}_\alpha$.

Consider the sparsity pattern of $|M_0|$:



All entries denoted $\times$ are non-zero.

From the sparsity pattern it can be deduced that all of the nodes are connected to nodes 5 and 6. To see this consider row 1, since $|m|_{1,5}$ and $|m|_{1,6}$ are non-zero, node 1 is connected to nodes 5 and 6. Similarly nodes 2 and 3 are connected to nodes 5 and 6. Node 4 is directly connected to node 1, so node 4 is also connected to nodes 5 and 6 (follow the paths linking node 4 to 1 and node 1 to nodes 5 and 6). Nodes 5 and 6 are connected to themselves via nodes 2 and 3. Since every node is directly connected to the third previous node (i.e. node 7 is directly connected to node 4) it follows by induction that all nodes are connected to nodes 5 and 6.

Furthermore, since nodes 2 and 5 are directly connected to each other it follows that all the nodes are also connected to node 2.

Now suppose α is such that $x_\gamma = 1$ for all $\gamma \in \mathcal{G}_\alpha$. Since $\{2, 5, 6\} \subset \mathcal{G}_\alpha$, it follows

that $x_2 = x_5 = x_6 = 1$. Therefore, from Lemma 3.3.7:

$$\lambda = \lambda x_2 = (|M_0|\boldsymbol{x})_2 = \sum_{j=1}^{3n} |m|_{2,j}x_j = |m|_{2,5}x_5 + |m|_{2,6}x_6 = |m|_{2,5} + |m|_{2,6} < 1. \quad (3.3.40)$$

Hence $\lambda < 1$ even if $x_\gamma = 1$ for all $\gamma \in \mathcal{G}_\alpha$. This completes the proof of part (i).

Part (ii) follows from Exercise 6.3.11 of [37] [$\rho(B) \leq \rho(A)$ if $|B| \leq A$].

Part (iii) follows directly. □

Stage II

This completes the first part of Theorem 3.3.6. Next it is shown that the ANF iteration applied to the linearised semiconductor system with small non-zero voltage converges, this is shown in the next lemma. For this lemma it is helpful to note that the one dimensional analogue of Theorem 2.3.5 states the following: There exists an $r > 0$ such that for all $\boldsymbol{\alpha} \in \mathcal{B}(\mathbf{0}, r)$:

IFT1 The finite element solution, $\boldsymbol{X}(\boldsymbol{\alpha})$, to (3.3.22) with scaled applied voltage $\boldsymbol{\alpha}$ is continuous with respect to $\boldsymbol{\alpha}$.

IFT2 $\mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha})$, the Fréchet derivative of (3.3.22) with respect to $\boldsymbol{X}$ is nonsingular.

IFT3 There exists an open set $\mathcal{D} \subset \mathbb{R}^{3n} \times \mathbb{R}^2$, containing $(\boldsymbol{X}(\mathbf{0}), \mathbf{0})$, such that for all $(\boldsymbol{Y}, \boldsymbol{\alpha}_1), (\boldsymbol{Y}, \boldsymbol{\alpha}_2), (\boldsymbol{Z}, \boldsymbol{\alpha}_1) \in \mathcal{D}$:

$$\|\mathbf{F}_X(\boldsymbol{Y}, \boldsymbol{\alpha}_1) - \mathbf{F}_X(\boldsymbol{Y}, \boldsymbol{\alpha}_2)\|_\infty \rightarrow 0 \text{ as } \|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_\infty \rightarrow 0, \quad (3.3.41)$$

$$\|\mathbf{F}_X(\boldsymbol{Y}, \boldsymbol{\alpha}_1) - \mathbf{F}_X(\boldsymbol{Z}, \boldsymbol{\alpha}_1)\|_\infty \rightarrow 0 \text{ as } \|\boldsymbol{Y} - \boldsymbol{Z}\|_\infty \rightarrow 0. \quad (3.3.42)$$

Lemma 3.3.9 *There exists an $r > 0$ such that for all $\boldsymbol{\alpha} \in \mathcal{B}(\mathbf{0}, r)$. the linear ANF iteration applied to the system*

$$A_\alpha \boldsymbol{X} = b_\alpha \quad (3.3.43)$$

(where $A_\alpha = \mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha})$ and $b_\alpha \in \mathbb{R}^{3n}$ is arbitrary) converges.

Proof The iteration matrix of the ANF iteration applied to (3.3.43) is defined by

$$M_\alpha = I - \sum_{\kappa=1}^{n} \rho_\kappa \left[\mathbf{F}_V^{-1}(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha})_\kappa\right] r_\kappa \mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha}). \quad (3.3.44)$$

In (3.3.44), $\mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}),\boldsymbol{\alpha})_\kappa := r_\kappa \mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}),\boldsymbol{\alpha})p_\kappa,\ \ \kappa = 1,2,\ldots,n.$

It is known from Lemma 3.3.8 that $\rho(M_0) < 1$. The aim of this lemma is to show that there exists an $r > 0$ such that for all $\boldsymbol{\alpha} \in \mathcal{B}(\mathbf{0},r)$ $\rho(M_\alpha) < 1$. This is shown by a perturbation argument with respect to $\boldsymbol{\alpha}$.

Since $\rho(M_0) < 1$ there exists, by Corollary 3.6 of [74, Chapter 2], a matrix norm $\|\cdot\|_*$ and a number $\varphi > 0$ such that

$$\|M_0\|_* \leq \varphi < 1.$$

If it can be shown that $\|M_\alpha\|_* < 1$ it follows from Theorem 3.4 of [74, Chapter 2] that $\rho(M_\alpha) < 1$, proving the Lemma.

$$\begin{aligned} \|M_\alpha\|_* &\leq \|M_\alpha - M_0\|_* + \|M_0\|_* \\ &\leq \|M_\alpha - M_0\|_* + \varphi. \end{aligned} \tag{3.3.45}$$

If $\|M_\alpha - M_0\|_* \to 0$ as $\|\boldsymbol{\alpha}\|_\infty \to 0$, then, for $\boldsymbol{\alpha}$ sufficiently small in norm, it follows that $\|M_\alpha\|_* < 1$. To show the continuity of M_α, with respect to $\boldsymbol{\alpha}$, consider:

$$\begin{aligned} \|M_\alpha - M_0\|_* &= \| \left[I - \sum_{\kappa=1}^{n} p_\kappa \mathbf{F}_X^{-1}(\boldsymbol{X}(\boldsymbol{\alpha}),\boldsymbol{\alpha})_\kappa r_\kappa \mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}),\boldsymbol{\alpha}) \right] - \\ &\qquad \left[I - \sum_{\kappa=1}^{n} p_\kappa \mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}),\mathbf{0})_\kappa r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\mathbf{0}) \right] \|_* \\ &= \| \sum_{\kappa=1}^{n} p_\kappa \mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}),\mathbf{0})_\kappa r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}).\mathbf{0}) - \\ &\qquad \sum_{\kappa=1}^{n} p_\kappa \mathbf{F}_X^{-1}(\boldsymbol{X}(\boldsymbol{\alpha}),\boldsymbol{\alpha})_\kappa r_\kappa \mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}),\boldsymbol{\alpha})\|_*. \end{aligned}$$

From the triangle inequality it follows that:

$$\begin{aligned} \|M_\alpha - M_0\|_* = &\| \sum_{\kappa=1}^{n} p_\kappa \left[\mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}).\mathbf{0})_\kappa - \mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}).\boldsymbol{\alpha})_\kappa \right] r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}).\mathbf{0})\|_* + \\ &\| \sum_{\kappa=1}^{n} p_\kappa \left[\mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}).\boldsymbol{\alpha})_\kappa - \mathbf{F}_X^{-1}(\boldsymbol{X}(\boldsymbol{\alpha}).\boldsymbol{\alpha})_\kappa \right] r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}).\mathbf{0})\|_* + \end{aligned}$$

$$\| \sum_{\kappa=1}^{n} p_\kappa \mathbf{F}_X^{-1}(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha})_\kappa r_\kappa \left[\mathbf{F}_X(\boldsymbol{X}(\mathbf{0}), \mathbf{0}) - \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha})\right] \|_* \; +$$
$$\| \sum_{\kappa=1}^{n} p_\kappa \mathbf{F}_X^{-1}(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha})_\kappa r_\kappa \left[\mathbf{F}_X(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha}) - \mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha})\right] \|_*. \quad (3.3.46)$$

From **[IFT1]**-**[IFT3]** (and the equivalence of all norms in a finite dimensional space) the last two terms of (3.3.46) tend to zero as $\boldsymbol{\alpha}$ decreases in norm. For $\|\boldsymbol{\alpha}\|_\infty$ sufficiently small it follows from **[IFT1]**-**[IFT3]** that:

$$\begin{aligned}
&\|\mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}), \mathbf{0})_\kappa - \mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha})_\kappa\|_* \\
&\quad = \|\mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}), \mathbf{0})_\kappa \left[\mathbf{F}_X(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha})_\kappa - \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}), \mathbf{0})_\kappa\right] \mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha})_\kappa\|_* \\
&\quad = \|\mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}), \mathbf{0})_\kappa r_\kappa \left[\mathbf{F}_X(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha}) - \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}), \mathbf{0})\right] p_\kappa \mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha})_\kappa\|_* \\
&\qquad \to 0 \;\text{ as } \|\boldsymbol{\alpha}\|_\infty \to 0.
\end{aligned}$$

Therefore the first term of (3.3.46) tend to zero as $\boldsymbol{\alpha}$ decreases in norm. Also

$$\begin{aligned}
&\|\mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha})_\kappa - \mathbf{F}_X^{-1}(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha})_\kappa\|_* \\
&\quad = \|\mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha})_\kappa \left[\mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha})_\kappa - \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha})_\kappa\right] \mathbf{F}_X^{-1}(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha})_\kappa\|_* \\
&\quad = \|\mathbf{F}_X^{-1}(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha})_\kappa r_\kappa \left[\mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) - \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}), \boldsymbol{\alpha})\right] p_\kappa \mathbf{F}_X^{-1}(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha})_\kappa\|_* \\
&\qquad \to 0 \;\text{ as } \|\boldsymbol{X}(\boldsymbol{\alpha}) - \boldsymbol{X}(\mathbf{0})\|_\infty \to 0.
\end{aligned}$$

Since $\boldsymbol{X}(\boldsymbol{\alpha})$ is continuous in $\boldsymbol{\alpha}$, it follows that the second term in (3.3.46) tend to zero as $\boldsymbol{\alpha}$ decreases in norm. This shows that M_α is continuous in $\boldsymbol{\alpha}$.

Since M_α is continuous in $\boldsymbol{\alpha}$, for sufficiently small $\boldsymbol{\alpha}$ in norm, it follows from (3.3.45) that there exists an $r > 0$ such that for all $\boldsymbol{\alpha} \in \mathcal{B}(\mathbf{0}, r)$:

$$\|M_\alpha\|_* < 1.$$

Completing the proof. □

The above lemma shows that the linear ANF iteration applied to the linearised semiconductor system (3.3.22) converges, for small applied voltage. Next we show that the nonlinear ANF iteration applied to the semiconductor system converges:

Stage III

Here we use Theorem 3.3.2 to prove Theorem 3.3.6. For this it is necessary to show that the assumptions **[H1]**-**[H5]** of the Dryja-Hackbusch hold for the system. This is the purpose of the next lemma:

Lemma 3.3.10 *There exist $r > 0$ such that for all $\boldsymbol{\alpha} \in \mathcal{B}(\mathbf{0}, r)$,* **[H1]**-**[H5]** *of Section 3.3.1 hold for the system:*

$$\mathbf{F}(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = \mathbf{0},$$

where $\mathbf{F}$ *is given by (3.3.22).*

Proof For α sufficiently small in norm, **[H1]**-**[H3]** follow from the one dimensional analogue of the application of the Implicit Function Theorem to the semiconductor equations as discussed in Theorem 2.3.5 of Chapter 2. The Implicit Function Theorem requires that there exists a solution, $\boldsymbol{X}(\mathbf{0})$, to (3.3.22) when $\boldsymbol{\alpha} = \mathbf{0}$ - it is easy to check that $\boldsymbol{X}(\mathbf{0})$ given by (3.3.23) is a solution to (3.3.22) (the existence and uniqueness of Ψ_0, satisfying (3.3.24) is given in Theorem 3.3 of [23]).

It follows from the application of the Implicit Function Theorem that the Fréchet derivative of $\mathbf{F}$, $\mathbf{F}_X$, exists and is continuous at the solution to (3.3.22) for all sufficiently small scaled applied voltages. With this in mind assumption **[H4]** follows immediately by taking

$$DF_\alpha(x', x'') = \int_0^1 \mathbf{F}_X(x' + t(x'' - x'), \boldsymbol{\alpha}) dt.$$

[The second point of assumption **[H4]** follows since $\mathbf{F}_X(\cdot, \boldsymbol{\alpha})$ is continuous at $\boldsymbol{X}(\boldsymbol{\alpha})$, for sufficiently small $\boldsymbol{\alpha}$].

Finally, it remains to verify, for **[H5]**, that $r_\kappa \mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) p_\kappa : X_\kappa \to X_k$ is invertible, $\kappa = 1, 2, \ldots, n$. First it is shown that $r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}), \mathbf{0}) p_\kappa$ is invertible.

Using the notation of Remark 3.3.5:

$$r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\mathbf{0})p_\kappa = \begin{bmatrix} \frac{\lambda^2}{h_\kappa} + \frac{\lambda^2}{h_{\kappa+1}} + & -e_\kappa(\boldsymbol{\Psi}_0) & -e_\kappa(-\boldsymbol{\Psi}_0) \\ e_\kappa(\boldsymbol{\Psi}_0) + e_\kappa(-\boldsymbol{\Psi}_0) & & \\ 0 & a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0) + & -\rho_v h_\kappa(\boldsymbol{\Psi}_0) \\ & \rho_v h_\kappa(\boldsymbol{\Psi}_0) & \\ 0 & -\rho_w h_\kappa(\boldsymbol{\Psi}_0) & a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0) + \\ & & \rho_w h_\kappa(\boldsymbol{\Psi}_0) \end{bmatrix}. \tag{3.3.47}$$

The $(1,1)$th entry of (3.3.47) is never zero and the $(1,1)$th cofactor is:

$$\begin{aligned} &\left\{\left[a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0) + \rho_v h_\kappa(\boldsymbol{\Psi}_0)\right]\left[a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0) + \rho_w h_\kappa(\boldsymbol{\Psi}_0)\right] - \rho_v\rho_w h_\kappa^2(\boldsymbol{\Psi}_0)\right\} \\ &\quad = \left\{\left[a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0)\right]\left[a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0)\right] + \right. \\ &\qquad \left. \rho_v h_\kappa(\boldsymbol{\Psi}_0)\left[a_\kappa(-\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(-\tilde{\boldsymbol{\Psi}}_0)\right] + \rho_w h_\kappa(\boldsymbol{\Psi}_0)\left[a_\kappa(\tilde{\boldsymbol{\Psi}}_0) + a_{\kappa+1}(\tilde{\boldsymbol{\Psi}}_0)\right]\right\} \end{aligned}$$

which is also never equal to zero. Thus $r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\mathbf{0})p_\kappa$ is invertible, $\kappa = 1,2,\ldots,n$.

By the application of the Implicit Function Theorem (recall (3.3.41) and (3.3.42)) $\mathbf{F}_X(\boldsymbol{X},\boldsymbol{\alpha})$ is continuous with respect to $\boldsymbol{X}$ and $\boldsymbol{\alpha}$. Thus $r_\kappa \mathbf{F}_X(\boldsymbol{X},\boldsymbol{\alpha})p_\kappa$ is also continuous with respect to $\boldsymbol{X}$ and $\boldsymbol{\alpha}$, since for example:

$$\begin{aligned} \|r_\kappa \mathbf{F}_X(\boldsymbol{X},\boldsymbol{\alpha})p_\kappa - r_\kappa \mathbf{F}_X(\boldsymbol{Y},\boldsymbol{\alpha})p_\kappa\|_\infty &= \|r_\kappa\left[\mathbf{F}_X(\boldsymbol{X},\boldsymbol{\alpha}) - \mathbf{F}_X(\boldsymbol{Y},\boldsymbol{\alpha})\right]p_\kappa\|_\infty \\ &\to 0 \text{ as } \|\boldsymbol{X} - \boldsymbol{Y}\|_\infty \to 0. \end{aligned}$$

Since $\boldsymbol{X}(\boldsymbol{\alpha})$ is continuous in $\boldsymbol{\alpha}$, it follows that, as $\|\boldsymbol{\alpha}\|_\infty \to 0$:

$$\begin{aligned} &\| \left(r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\mathbf{0})p_\kappa\right)^{-1}\left[r_\kappa \mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}),\boldsymbol{\alpha})p_\kappa - r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\mathbf{0})p_\kappa\right] \|_\infty \\ &\quad \le \| \left(r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\mathbf{0})p_\kappa\right)^{-1}\left[r_\kappa \mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}),\boldsymbol{\alpha})p_\kappa - r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\boldsymbol{\alpha})p_\kappa\right] \|_\infty + \\ &\qquad \| \left(r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\mathbf{0})p_\kappa\right)^{-1}\left[r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\boldsymbol{\alpha})p_\kappa - r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\mathbf{0})p_\kappa\right] \|_\infty \\ &\qquad \to 0. \end{aligned}$$

Since $r_\kappa \mathbf{F}_X(\boldsymbol{X}(\mathbf{0}),\mathbf{0})p_\kappa$ is invertible, it follows from Theorem A.2.7, that for $\boldsymbol{\alpha}$ sufficiently small, $r_\kappa \mathbf{F}_X(\boldsymbol{X}(\boldsymbol{\alpha}),\boldsymbol{\alpha})p_\kappa$ is invertible. Proving **[H5]**. □

It is now possible to prove Theorem 3.3.6, this states that the nonlinear ANF iteration applied to the nonlinear system (3.3.22) converges for sufficiently small scaled applied voltage.

Proof of Theorem 3.3.6

Since the spectral radius of the ANF iteration matrix is less than one for small applied voltage, the linear ANF iteration applied to the linearisation of (3.3.22) converges (see [54]).

Finally, since assumptions **[H1]**-**[H5]** hold for the ANF method applied to (3.3.22), it follows from Theorem 3.3.2 that the nonlinear ANF method applied to (3.3.22) converges for sufficiently small scaled applied voltages. □

Remark 3.3.11 *Since completing this thesis it has come to our attention that it is possible to prove Lemma 3.3.8 (the proof that the linear ANF iteration with zero applied voltage converges) using the M-matrix theory of Hackbush [37, Chapter 6]. This result allows us to extend the convergence of the ANF iteration to the two dimensional case. The proof takes the following form:*

In Chapter 2 it was shown that the linearisation of the two dimensional semiconductor system with zero applied voltage, (2.3.31), was essentially diagonally dominant (Theorem 2.3.5). It is also easy to see from Lemma 2.3.3 and (2.3.31) that this matrix has positive diagonal terms and negative off diagonal terms and is therefore an M-matrix ([37, Theorem 6.4.4]).

Recalling from Section 3.2.1 that the linear ANF iteration is a type of block Jacobi iteration it can be deduced from [37, Theorem 6.1.1] that the linear ANF iteration applied to the two dimensional semiconductor problem with no applied voltage across the device converges. Analagous arguments to those used in the one dimensional ANF proof show that the convergence of the linearised iteration can be extended to the case of sufficently small applied voltage. Finally an application of [29] shows that the nonlinear Jacobi-ANF method converges.

3.4 Numerical Results for the ANF Method

In this section the convergence and performance of the ANF method is compared with the standard Gummel and Newton Gauss-Seidel methods for solving the drift-diffusion

equations. It is shown that the ANF method applied to the semiconductor system (3.3.22) converges for a larger range of applied voltages than the other two methods.

The ANF method is applied to the mass lumped finite element discretisation of the semiconductor equations (3.3.14)-(3.3.16) on a uniform mesh with n interior mesh points.

The ANF method implemented is based on the alternative blocking described in Section 3.2.1. For an applied voltage of V_l at the left hand boundary ($x = 0$ in the model) and V_r at the right hand boundary ($x = 1$), the iteration is:

1 Make an initial guess $\boldsymbol{X}_0 = \left[\boldsymbol{\Psi}^{0^T}, \boldsymbol{V}^{0^T}, \boldsymbol{W}^{0^T}\right]^T \in \mathbb{R}^{3n}$ to the solution of the system (3.3.22). Set k = 0.

2 For $\kappa = 1, 2, \ldots, n$, find

$$\left(\left(\boldsymbol{\Psi}^{k+1}\right)_\kappa, \left(\boldsymbol{V}^{k+1}\right)_\kappa, \left(\boldsymbol{W}^{k+1}\right)_\kappa\right)$$

such that (for $\boldsymbol{z} = \boldsymbol{\Psi}, \boldsymbol{V}, \boldsymbol{W}$)

$$\left(\tilde{\boldsymbol{z}}^{k+1,\kappa}\right)_l = \begin{cases} \left(\boldsymbol{z}^k\right)_l & l \neq \kappa \\ \left(\boldsymbol{z}^{k+1}\right)_l & l = \kappa \end{cases} \qquad l = 1, 2, \ldots, n$$

solves

$$r_\kappa \mathbf{F}\left(\left[\tilde{\boldsymbol{\Psi}}^{k+1,\kappa^T}, \tilde{\boldsymbol{V}}^{k+1,\kappa^T}, \tilde{\boldsymbol{W}}^{k+1,\kappa^T}\right]^T, \alpha_0, \alpha_1\right) = 0.$$

In the above $\alpha_0 = \frac{V_l}{U_T}$, $\alpha_1 = \frac{V_r}{U_T}$, $\mathbf{F}$ is given by (3.3.22) and r_κ is given by (3.3.27).

3 Set $\boldsymbol{X}_{k+1} = \left[\boldsymbol{\Psi}^{k+1^T}, \boldsymbol{V}^{k+1^T}, \boldsymbol{W}^{k+1^T}\right]^T$ to be the vector whose values were calculated in 2. Set k = k+1.

4 If the difference between $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k+1}$ in the 2-norm is less than 5×10^{-5} then stop, otherwise return to step 2.

The nonlinear system in Step 2 is solved by Newton's method. The initial guess for ψ, v and w is the doping profile scaled to match the relevant Dirichlet boundary conditions.

The convergence results for the ANF method applied to the PN diode problem are presented in Table 3.2. The method was tested for a range of applied voltages and for a range of uniform finite element meshes. Comparing these results with those of

Number of interior mesh points	Voltage applied at 0	Voltage applied at 1	Number of outer loops for convergence
9	0	0.1	42
9	0	0.2	47
9	0	0.5	54
9	0	1	61
9	0	1.4	Diverges
9	0.1	0	43
9	0.2	0	Diverges
21	0	0.1	156
21	0	0.5	228
21	0	1	267
21	0	1.4	Diverges

Table 3.2: Convergence of the ANF method applied to a simple one dimensional PN diode with doping profile equal to -1 on $[0, 1/2)$ and +1 on $(1/2, 1]$. The ANF method breaks down when an voltage of greater than one volt is applied across the device.

Gummel's method (described in Section 3.3.3) and the Newton method (described in Section 2.4) on the same problem it is seen that the ANF method converges for a larger range of voltages in reverse bias and will converge for small forward bias. However, as the ANF method has to solve a large number of 3×3 nonlinear systems the method is slow in comparison with Gummel's or Newton's method. It is believed that the parallel implementation of the ANF method will be far superior to the Gummel or Newton methods both in terms of speed and size of convergence ball.

The ANF method breaks down when the applied voltage is increased above 1 volt; this seems to be due to the inner Newton iteration failing to converge. To combat this globally convergent Newton methods were investigated, details of such methods can be found in [28]. We used a hookstep method taken from a set of programs based on the outlines in [28] and coded by R. Behrens. We implemented the globally convergent Newton method only for the semiconductor system (3.3.14)-(3.3.16) with the generation/recombination term, r given by (3.3.17), set to zero. The globally convergent Newton method did increase the range of applied voltages we were able to solve for by 0.4 volts.

Chapter 4

Iterated Defect Correction for Irregular Semilinear Problems

In this chapter we consider semilinear scalar equations. This is useful preparation for solving semiconductor equations, since many iterative methods for the drift-diffusion system consist of solving sequences of semilinear equations (for example the version of Gummel's method discussed in Remark 3.3.4). In addition, when there is no applied voltage across the device the whole drift-diffusion system reduces to the electrostatic potential equation:

$$-\lambda^2 \Delta \psi + 2\delta^2 \sinh \psi - d = 0,$$

where λ^2 and δ^2 are both small and d is the doping profile.

In this chapter we show that there exists a finite element solution to a general semilinear problem posed on a domain with a polygonal boundary subject to mixed boundary conditions. We prove *a priori* error estimates and introduce an efficient multilevel method for accurate solution of the discretised problem.

A review of the originality of the results in this chapter is given in Section 4.3.1.

4.1 The Problem and Basic Definitions

Let Ω be a bounded domain with polygonal boundary $\partial\Omega$. Assume $\partial\Omega = \cup_{j=1}^{\nu} \partial\Omega_j$ where $\partial\Omega_j$, $j = 1, \ldots, \nu$ are consecutive straight line segments of $\partial\Omega$, numbered as $\partial\Omega$ is traversed in an anti-clockwise direction. Consider the solution of the following

equation:

$$-\Delta u(\boldsymbol{x}) + f(u(\boldsymbol{x}), \boldsymbol{x}) = 0 \text{ in } \Omega, \tag{4.1.1}$$

subject to the boundary conditions:

$$u = g \text{ on } \partial\Omega_D, \tag{4.1.2}$$

$$\frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega_N. \tag{4.1.3}$$

Where the Dirichlet and Neumann boundaries $\partial\Omega_D$ and $\partial\Omega_N$ are assumed to form a partition of $\partial\Omega$: $\{\partial\Omega_j : j = 1, \ldots, \nu\}$.

Identify $\partial\Omega_{\nu+1}$ with $\partial\Omega_1$ and set $\boldsymbol{x}_j = \overline{\partial\Omega_j} \cap \overline{\partial\Omega_{j+1}}$ for each $j = 1, \ldots, \nu$. Let $\omega_j \in (0, 2\pi)$ denote the angle (internal to Ω) between the segments $\partial\Omega_j$ and $\partial\Omega_{j+1}$ at $\boldsymbol{x}_j$. When $\partial\Omega_j$ and $\partial\Omega_{j+1}$ both belong to either $\partial\Omega_D$ or $\partial\Omega_N$ it will only be necessary to consider $\boldsymbol{x}_j$ to be a *corner point* (i.e. $\omega_j \neq \pi$). Otherwise the solution is smooth near $\boldsymbol{x}_j$. If $\boldsymbol{x}_j$ is a *collision point* (i.e. $\boldsymbol{x}_j \in \overline{\partial\Omega_D} \cap \overline{\partial\Omega_N}$) then it is necessary to consider all $\omega_j \in (0, 2\pi)$.

If any $\boldsymbol{x}_j$ is a collision point and/or if $\omega_j > \pi$ for any j, then the solution u of the problem (4.1.1)-(4.1.3) will not be in the Sobolev space H^2 near $\boldsymbol{x}_j$, but rather will have a singularity of the form $|\boldsymbol{x} - \boldsymbol{x}_j|^{\alpha_j}$, where $\alpha_j < 1$ depends on ω_j. This is proved in [35] for (4.1.1)-(4.1.3) in the case $f \equiv 0$. The result is established in this thesis for the case of quite general nonlinear f.

At this stage of the thesis the following quite general assumptions are made

(A1) $f : \mathbb{R} \times \Omega \to \mathbb{R}$ has the property that for all $\boldsymbol{x} \in \Omega$, $f(\cdot, \boldsymbol{x}) \in C^2(\mathbb{R})$ and if $u \in C(\Omega)$ then the function $\boldsymbol{x} \to f(u(\boldsymbol{x}), \boldsymbol{x})$ is in $\mathrm{L}_\infty(\Omega)$.

(A2) $g \in \mathrm{H}^{\frac{3}{2}}(\partial\Omega_D)$.

(A3) (4.1.1)-(4.1.3) has a solution $u_0 \in \mathrm{L}_\infty(\Omega)$ with the property that $\Delta u_0 \in \mathrm{L}_\infty$.

Remark 4.1.1 i **(A1)** *allows the case where f is a smooth, but not globally bounded function of u.*

ii **(A2)** *ensures that g has an extension $g_e \in H^2$, with $g_e|_{\partial\Omega_D} = g$.*

iii *Many sufficient conditions for* **(A3)** *can be found in the literature (e.g. [41] and [64], although many published results are only for smooth boundaries).*

We use C and C' to denote generic positive constants. For convenience introduce the shorthand notation $f(u) := f(u(\boldsymbol{x}), \boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, $f'(u)$ and $f''(u)$ will be taken to mean the derivatives of f with respect to u.

To describe the solution of (4.1.1)-(4.1.3) it is necessary to refer to the fractional order Sobolev spaces $\mathrm{H}^s = \mathrm{H}^s(\Omega)$, $s \geq 0$, as defined in [35, Section 1.2]. As usual $|\cdot|_s$ and $\|\cdot\|_s$ denote, respectively, the semi-norm and norm in H^s. $|\cdot|_0 = \|\cdot\|_0$ denotes the usual L_2 norm on Ω and $\|\cdot\|_\infty$ denotes the uniform norm on Ω. For a subdomain L of Ω, $\|\cdot\|_{s,L}$ denotes the norm on $\mathrm{H}^s(L)$. If L is a subset of the boundary of Ω, $\mathrm{H}^s(L)$ is taken to be the trace space as defined in [35, Section 1.4].

It is also necessary to define two further spaces which are used for notational purposes in the finite element analysis. Define H^2_W to be a weighted H^2 space with a weight function W which decays sufficiently quickly near the points $\boldsymbol{x}_j$, such that the weak solution of (4.1.1)-(4.1.3) is a member of H^2_W. Denote the norm of H^2_W by $\|\cdot\|_{\mathrm{H}^2_W}$. Let $\mathcal{C}^2_W$ be an analogous weighted $\mathcal{C}^2$-space, with norm $\|\cdot\|_{\mathcal{C}^2_W}$. The precise form of the weight function is not needed here, we simply need the fact that a suitable weight exists as in [60].

Define $\mathcal{V} = \mathrm{H}^1$ and for any Dirichlet data, $d \in \mathrm{H}^{\frac{3}{2}}(\partial\Omega_D)$:

$$\mathcal{V}_d := \{v \in \mathcal{V} : \ v|_{\partial\Omega_D} = d\} .$$

The following lemma is fundamental in our quest to fully describe the solution of (4.1.1)-(4.1.3).

Lemma 4.1.2 *Suppose $b \in L_2$ and u is the unique element of $\mathcal{V}_g$ which satisfies*

$$(\nabla u, \nabla v) = (b, v). \quad \textit{for all } v \in \mathcal{V}_0. \tag{4.1.4}$$

Then there exists an $\alpha \in (1/4, 1]$, depending only on the angles $\{\omega_j\}$, and a constant C which is independent of b, g_e and u, such that:

$$\|u\|_{1+\alpha} \leq C \{\|b\|_0 + \|g_e\|_2\} . \tag{4.1.5}$$

Proof The lemma is proved by combining several results from Grisvard [35].

Set $B = \|b\|_0 + \|\Delta g_e\|_0 + \|\frac{\partial g_e}{\partial n}|_{\partial\Omega_D}\|_{\frac{1}{2},\partial\Omega_D}$.

By [35, Theorem 1.4.6] $B \leq C\{\|b\|_0 + \|g_e\|_2\}$.

Define $\tilde{u} := u - g_e$. Then $\tilde{u} \in \mathcal{V}_0$ and using (4.1.4) and Green's Theorem yields

$$\begin{aligned}(\nabla\tilde{u}, \nabla v) &= (b, v) + (\Delta g_e, v) - \left(\frac{\partial g_e}{\partial n}, v\right)_{\partial\Omega_D} \\ &=: L(v)\end{aligned}$$

for all $v \in \mathcal{V}_0$.

Since

$$\begin{aligned}\left(\frac{\partial g_e}{\partial n}, v\right)_{\partial\Omega_D} &\leq \|\frac{\partial g_e}{\partial n}\|_{\frac{1}{2},\partial\Omega_D}\|v\|_{-\frac{1}{2},\partial\Omega_D} \\ &\leq \|\frac{\partial g_e}{\partial n}\|_{\frac{1}{2},\partial\Omega_D}\|v\|_0,\end{aligned}$$

L is a linear function on L_2 with norm bounded by B. Thus it follows from Section 2.5.2 and Theorem 2.5.2 of [35] that for each j, there exist singular functions $S_{j,m}$ ($m = 1, \ldots, n_j$, n_j finite, for each j) and scalars c_j such that

$$u_s = \sum_{j=1}^{\nu}\sum_{m=1}^{n_j} c_j S_{j,m} \tag{4.1.6}$$

such that

$$u_R := \tilde{u} - u_s$$

satisfies

$$\|u_R\|_2 \leq CB \tag{4.1.7}$$

for some constant C. Each function $S_{j,m}(\boldsymbol{x})$ is smooth except for a singularity which is no worse than $|\boldsymbol{x} - \boldsymbol{x}_j|^{\alpha_j}$ as $\boldsymbol{x} \to \boldsymbol{x}_j$, where $\alpha_j > 1/4$ for each j.

For each j it follows from [35, Theorem 1.2.18] that

$$S_{j,m} \in \mathrm{H}^{1+\alpha_j}, \quad m = 1, \ldots, n_j.$$

Also from [35, Theorem 2.5.2] the coefficients c_j in (4.1.6) satisfy

$$|c_j| \leq CB.$$

With $\alpha = \min\{\alpha_j\}$ it follows, from (4.1.6), that

$$\begin{aligned} \|u_s\|_{1+\alpha} &\leq CB \sum_{j=1}^{\nu} \sum_{m=1}^{n_j} \|S_{j,m}\|_{1+\alpha} \\ &\leq CB, \end{aligned} \tag{4.1.8}$$

with C independent of u. Then (4.1.8) together with (4.1.7) implies

$$\begin{aligned} \|\tilde{u}\|_{1+\alpha} &\leq \|u_s\|_{1+\alpha} + \|u_R\|_{1+\alpha} \\ &\leq CB, \end{aligned}$$

and since $u = \tilde{u} + g_e$ the result follows. □

For now on let α denote the number found in Lemma 4.1.2.

Corollary 4.1.3 *Let $u_0 \in L_\infty$ be the solution to (4.1.1)-(4.1.3) introduced in assumption* **(A3)**. *Then $u_0 \in H^{1+\alpha}$ and*

$$\|u_0\|_{1+\alpha} \leq C \left\{ \|f(u_0)\|_0 + \|g_e\|_2 \right\}. \tag{4.1.9}$$

Proof For $u_0 \in \mathrm{L}_\infty$ it follows from **(A1)**, **(A3)** that $f(u_0) \in \mathrm{L}_\infty \subset \mathrm{L}_2$, so the result follows from Lemma 4.1.2 (since the strong solution of (4.1.1)-(4.1.3) is also a weak solution). □

In the next section we will consider the solution of (4.1.1)-(4.1.3) which is a zero of $\mathrm{F} : \mathcal{V}_g \to (\mathcal{V}_0)'$ defined by:

$$(\mathrm{F}(u), v) := (\nabla u, \nabla v) + (f(u), v) = 0, \quad u \in \mathcal{V}_g, v \in \mathcal{V}_0. \tag{4.1.10}$$

This has linearisation

$$\left(\mathrm{F}'(u)v, w\right) = (\nabla v, \nabla w) + (f'(u)v, w), \quad u \in \mathcal{V}_g, v, w \in \mathcal{V}_0. \tag{4.1.11}$$

At this stage make the following additional assumptions on our problem:

(A4) $\mathrm{F}'(u_0) : \mathcal{V}_0 \to (\mathcal{V}_0)'$ is bijective and hence, by Banach's isomorphism theorem, for all $w \in \mathcal{V}_0$

$$\|\mathrm{F}'(u_0)w\|_{(\mathcal{V}_0)'} \geq C\|w\|_1$$

Remark 4.1.4 *Assumption* **(A4)** *ensures that for all* $b \in L_2$, *there exists a unique* $w \in \mathcal{V}_0$ *such that*

$$\left(F'(u_0)w, v\right) = (b, v), \quad \textit{for all } v \in \mathcal{V}_0.$$

From Lemma 4.1.2 and **(A4)** *it follows that* $w \in H^{1+\alpha}$ *and :*

$$\|w\|_{1+\alpha} \leq C\|b\|_0.$$

4.2 The Finite Element System

For weak solutions of problems of the form (4.1.1)-(4.1.3), the finite element method combined with quasi-uniform mesh refinement will only yield suboptimal convergence rates due to the singularities appearing at the $\boldsymbol{x}_j$. However it is known that mesh grading will restore optimal convergence (see, for example [60, Section 7]). Surprisingly there seems to be no literature on the basic stability/convergence theory of the finite element method applied to (4.1.1)-(4.1.3) under the general conditions considered here. Hence we give such a theory in this section.

In this section it is assumed that there exists a sequence of meshes $\mathcal{T}_h$ with the following properties:

(M1) The meshes are shape regular (non-degenerate) in the sense of Ciarlet [20].

(M2) The number of triangles in $\mathcal{T}_h$ is of order h^{-2} as $h \to 0$, where h denotes the maximum diameter of the triangles in $\mathcal{T}_h$.

(M3) The interpolant $\Pi_h u_0$ to u_0 at the mesh points of the mesh satisfies

$$\|u_0 - \Pi_h u_0\|_1 \leq Ch \|u_0\|_{\mathrm{H}^2_{\mathrm{W}}}, \tag{4.2.12}$$

$$\|u_0 - \Pi_h u_0\|_0 \leq Ch^2 \|u_0\|_{\mathrm{H}^2_{\mathrm{W}}}. \tag{4.2.13}$$

$$\|u_0 - \Pi_h u_0\|_\infty \leq Ch^2 \|u_0\|_{\mathcal{C}^2_{\mathrm{W}}}. \tag{4.2.14}$$

Where the weighted norms are chosen as indicated in Section 4.1.

(M4)

$$h\left(\log\left(\frac{1}{h}\right)\right)^{\frac{1}{2}} \to 0 \text{ monotonically as } h \to 0.$$

where $\underline{h}$ denotes the minimum diameter of the triangles in $\mathcal{T}_h$.

Remark 4.2.1 (M4) *is a very weak condition which says that the minimum diameter of the triangles should not become too small compared to the maximum diameter.*

An example of a sequence of *a priori* defined meshes which satisfy **(M1)-(M4)** is given in [60, Theorem 1.7.2]. This construction is given only for the case of a single singular point $\boldsymbol{x}_j$, but the extension to many $\boldsymbol{x}_j$ is straightforward in principle. More generally adaptive procedures aim at satisfying **(M1)-(M4)** by *a posteriori* error estimation techniques. In this chapter it is assumed, for the theory, that **(M1)-(M4)** are satisfied.

Since it has been assumed that the triangulation, $\mathcal{T}_h$, is non-degenerate, the 'quasi-interpolant' $\tilde{w}$ of $w \in \mathrm{H}^{1+\alpha}$, [61, Theorem 4.1], satisfies

$$\|w - \tilde{w}\|_1 \leq Ch^{\alpha} \|w\|_{1+\alpha}. \tag{4.2.15}$$

This will be used in Lemma 4.2.9.

It remains to define the finite element approximation to the weak solution of (4.1.1)-(4.1.3). First define the piecewise linear finite element space:

$$\mathcal{V}_h := \left\{v \in \mathrm{H}^1 : v \text{ is continuous on } \Omega,\ v|_T \text{ is linear for all triangles } T \in \mathcal{T}_h\right\}$$

and for $d \in \mathrm{H}^{\frac{3}{2}}(\partial\Omega_D)$:

$$\mathcal{V}_{h,d} := \left\{v \in \mathcal{V}_h :\ v(\boldsymbol{x}) = d(\boldsymbol{x}) \text{ for all mesh points } \boldsymbol{x} \in \partial\Omega_D\right\}.$$

We consider the discrete problem of finding $u_h \in \mathcal{V}_{h,g}$ such that

$$\mathrm{F}_h(u_h) = 0 \ \text{ in } (\mathcal{V}_{h,0})'. \tag{4.2.16}$$

where $\mathrm{F}_h : \mathcal{V}_{h,g} \to \mathcal{V}_{h,0}$ is defined by:

$$(\mathrm{F}_h(u_h), v_h) = (\nabla u_h, \nabla v_h) + (f(u_h), v_h), \quad v_h \in \mathcal{V}_{h,0}. \tag{4.2.17}$$

In this subsection we show that there exists a unique u_h satisfying (4.2.16) and prove an optimal error estimate in the H^1-norm. Alternative proofs of the existence of finite

element solutions to (4.1.1)-(4.1.3) can be found in various references, for example in [25]. Here we give a proof using a more elementary starting point.

First recall the following well known lemma, which will be used in the proof of Lemma 4.2.3:

Lemma 4.2.2 [5, Chapter 2, Proposition 1.1]
Suppose $A : X \to Y$ is a bounded linear invertible map between Banach spaces X and Y. If $B : X \to Y$ is linear and

$$\| A - B \|_{X \to Y} \leq \frac{1}{\| A^{-1} \|_{Y \to X}},$$

then B is invertible.

The following key lemma studies the behaviour of the linearised operator $\mathrm{F}'(u)$, for u near u_0.

Lemma 4.2.3 *There exists $\delta > 0$ and $C > 0$ such that, for all $u \in \mathcal{V}_g \cap L_\infty$ with $\| u_0 - u \|_\infty \leq \delta$:*

(i) *$F'(u) : \mathcal{V}_0 \to (\mathcal{V}_0)'$ is bijective and*

$$\| v \|_1 \leq C \| F'(u)v \|_{-1}, \quad v \in \mathcal{V}_0.$$

(ii) *$F'(u) : \mathcal{V}_{h,0} \to (\mathcal{V}_{h,0})'$ is bijective and*

$$\| v_h \|_1 \leq C \| F'(u)v_h \|_{(\mathcal{V}_{h,0})'}, \quad v_h \in \mathcal{V}_{h,0}.$$

Proof If $\| u_0 - u \|_\infty \leq \delta \leq 1$, then using assumption **(A1)**:

$$\| f'(u_0) - f'(u) \|_\infty \leq C \| u - u_0 \|_\infty \leq C\delta.$$

The constant C depends on u_0 but not on u. From this it follows that for $v, w \in \mathcal{V}_0$,

$$\begin{aligned} \left| \left(\left[\mathrm{F}'(u_0) - \mathrm{F}'(u) \right] v, w \right) \right| &= \left| \left(\left[f'(u_0) - f'(u) \right] v, w \right) \right| \\ &\leq \| f'(u_0) - f'(u) \|_\infty \| v \|_1 \| w \|_1 \\ &\leq C\delta \| v \|_1 \| w \|_1. \end{aligned}$$

(as usual $\| \cdot \|_1$ denotes the norm in H^1), which implies

$$\| F'(u_0) - F'(u) \|_{\mathcal{V}_0 \to (\mathcal{V}_0)'} \leq C\delta. \tag{4.2.18}$$

Hence taking δ small enough it follows from Lemma 4.2.2 that $F'(u)$ is invertible, thus $F'(u)$ is bijective. To complete the proof of part (i), use assumption **(A4)** and (4.2.18), to obtain:

$$\begin{aligned} \| v \|_1 &\leq C_1 \| F'(u_0)v \|_{-1} \\ &\leq C_1 \left\{ \| \left[F'(u_0) - F'(u) \right] v \|_{-1} + \| F'(u)v \|_{-1} \right\} \\ &\leq C_1 \left\{ C_2\delta \| v \|_1 + \| F'(u)v \|_{-1} \right\} \end{aligned}$$

Choosing $\delta < \frac{1}{2C_1C_2}$ gives the required estimate.

For part (ii), injectivity follows from part (i). To prove surjectivity, consider a $d \in (\mathcal{V}_{h,0})'$ and the problem of finding a $v_h \in \mathcal{V}_{h,0}$ such that:

$$\left(F'(u)v_h, w_h \right) = (d, w_h), \quad w_h \in \mathcal{V}_h$$

this problem reduces to n equations in n unknowns for the coefficients of v_h. So the surjectivity follows from the injectivity.

Finally, define the Ritz projection: $P_h : \mathcal{V}_0 \to \mathcal{V}_{h,0}$, by

$$(\nabla P_h w, \nabla v_h) = (\nabla w, \nabla v_h), \ w \in \mathcal{V}_0, \ v_h \in \mathcal{V}_{h,0}. \tag{4.2.19}$$

Using (4.2.19) it is straightforward to deduce that for any $w \in \mathcal{V}_0$:

$$\| P_h w \|_1 \leq C \| w \|_1 \tag{4.2.20}$$

and by a duality argument it follows that:

$$\| w - P_h w \|_0 \leq Ch^\alpha \| w \|_1, \quad w \in \mathcal{V}_0. \tag{4.2.21}$$

For all $w \in \mathcal{V}_0$, $v_h \in \mathcal{V}_{h,0}$ it then follows from (4.2.19) and (4.2.21) that :

$$(F'(u)v_h, P_h w) = (\nabla v_h, \nabla P_h w) + (f'(u)v_h, P_h w)$$

$$
\begin{aligned}
&= (\nabla v_h, \nabla w) + (f^{'}(u)v_h, P_h w) \\
&= (\mathrm{F}^{'}(u)v_h, w) - (f^{'}(u)v_h, w - P_h w) \\
&\geq (\mathrm{F}^{'}(u)v_h, w) - C\|v_h\|_0 \|w - P_h w\|_0 \\
&\geq (\mathrm{F}^{'}(u)v_h, w) - Ch^{\alpha} \parallel v_h \parallel_1 \parallel w \parallel_1 .
\end{aligned}
$$

Since $P_h w_h = w_h$, for $w_h \in \mathcal{V}_{h,0}$, the result follows from the estimate in part (i) of this lemma. □

If, in addition to **(A1)-(A4)**, we were to assume that $f^{'} \geq 0$, then there are relatively simple arguments for proving the existence and uniqueness of finite element solutions to (4.2.16) [see for example the type of argument used in [42]]. Due to the quite general assumptions imposed in this section it is necessary to use Brouwer's Fixed Point Theorem to prove the existence of a solution u_h of (4.2.16). The argument, which is essentially adapted from Xu [73], uses a fixed point map, which is defined with the help of the following lemma:

Lemma 4.2.4 *u_h satisfies equation (4.2.16) if and only if*

$$
F^{'}(u_0)u_h = F^{'}(u_0)u_0 + R(u_0, u_h), \tag{4.2.22}
$$

where $R(u_0, u_h) \in (\mathcal{V}_{h,0})^{'}$ is defined by

$$
R(u_0, u_h)v_h = \left(f(u_0) + f^{'}(u_0)\,[u_h - u_0] - f(u_h), v_h\right). \tag{4.2.23}
$$

Proof Recall that $\mathrm{F}(u_0) = 0$ in $(\mathcal{V}_0)^{'}$. Then (4.2.16) is equivalent to $\mathrm{F}(u_0) - \mathrm{F}_h(u_h) = 0$ in $(\mathcal{V}_{h,0})^{'}$, which is equivalent to

$$
\mathrm{F}^{'}(u_0)\,[u_h - u_0] = \mathrm{F}(u_0) + \mathrm{F}^{'}(u_0)\,[u_h - u_0] - \mathrm{F}_h(u_h). \quad \text{in } (\mathcal{V}_{h,0})^{'} .
$$

This may be rewritten as:

$$
\begin{aligned}
&(\nabla(u_h - u_0), \nabla v_h) + (f^{'}(u_0)(u_h - u_0), v_h) \\
&\quad = (\nabla u_0, \nabla v_h) + (f(u_0), v_h) + (\nabla(u_h - u_0), \nabla v_h) + (f^{'}(u_0)(u_h - u_0), v_h) \\
&\qquad - (\nabla u_h, \nabla v_h) - (f(u_h), v_h).
\end{aligned}
$$

This equality is equivalent to:

$$\begin{aligned}(\nabla u_h, \nabla v_h) + (f'(u_0)u_h, v_h) &= (\nabla u_0, \nabla v_h) + (f(u_0) + f'(u_0)u_h - f(u_h), v_h) \\ &= (\nabla u_0, \nabla v_h) + (f'(u_0)u_0, v_h) + \\ &\quad (f(u_0) + f'(u_0)[u_h - u_0] - f(u_h), v_h),\end{aligned}$$

which is true if and only if

$$\mathrm{F}'(u_0)u_h = \mathrm{F}'(u_0)u_0 + R(u_0, u_h), \quad \text{in } (\mathcal{V}_{h,0})'$$

where $R(u_0, u_h)$ is given by (4.2.23), as required. □

Lemma 4.2.4 leads to the following definition:

Definition 4.2.5 *Define the map $\Phi_h : \mathcal{V}_{h,0} \to \mathcal{V}_{h,0}$ as follows, for each $v_h \in \mathcal{V}_{h,0}$ require $\Phi_h(v_h)$ to satisfy the equation:*

$$F'(u_0)\left[\Phi_h(v_h) + \Pi_h g_e\right] := F'(u_0)u_0 + R(u_0, v_h + \Pi_h g_e), \quad \text{in } (\mathcal{V}_{h,0})'. \tag{4.2.24}$$

Φ_h is well defined by Lemma 4.2.3(ii). The existence of a solution, u_h, to (4.2.16) is guaranteed if it can be shown that Φ_h has a fixed point, v_h, for then $u_h = \Phi_h(v_h) + \Pi_h g_e$ solves (4.2.16). The existence of such a fixed point is proved by applying Brouwer's Fixed Point Theorem. First we state the theorem:

Theorem 4.2.6 (Brouwer's Fixed Point Theorem) [Theorem 8.1.1 of [40]] *Let C be a bounded closed convex non-empty subset of a finite dimensional normed vector space, and suppose G is a continuous function that maps C into C. Then G has a fixed point in C; i.e. there exists $w \in C$ such that*

$$G(w) = w$$

In order to apply Brouwer's Fixed Point Theorem it is is necessary to prove Φ_h is continuous from a bounded closed convex subset of $\mathcal{V}_{h,0}$ to the same set. To define this set the following definition is needed:

Definition 4.2.7 *For any $u \in \mathcal{V}_g$, define $Q_h u \in \mathcal{V}_{h,g}$ to be the solution of the problem*

$$F'(u_0)Q_h u = F'(u_0)u \quad in\,(\mathcal{V}_{h,0})'. \tag{4.2.25}$$

Remark 4.2.8 *1. Although we have defined $F'(u_0) : \mathcal{V}_0 \to (\mathcal{V}_0)'$ the definition is easily extended to include $F'(u_0) : \mathcal{V}_g \to (\mathcal{V}_0)'$.*

2. A unique solution to (4.2.25) is guaranteed by Lemma 4.2.3(ii) by solving the problem $F'(u_0)v_h = F'(u_0)(u - \Pi_h g_e)$ in $(\mathcal{V}_{h,0})'$ for $v_h \in \mathcal{V}_{h,0}$ then setting $Q_h u = v_h + \Pi_h g_e$.

In particular, with $u = u_0$ define the following ball in $\mathcal{V}_{h,0}$ centred at $Q_h u_0 - \Pi_h g_e$:

$$\mathcal{B}_h = \left\{ v_h \in \mathcal{V}_{h,0} : \| v_h + \Pi_h g_e - Q_h u_0 \|_1 \leq h \left[\| u_0 \|_{H^2_W} + \| g_e \|_2 \right] \right\} \tag{4.2.26}$$

To prove the mapping properties of Φ_h the following optimal convergence property of $Q_h u_0$ is needed:

Lemma 4.2.9 *For small enough h:*

$$\| u_0 - Q_h u_0 \|_1 \leq Ch \left[\| u_0 \|_{H^2_W} + \| g_e \|_2 \right] \tag{4.2.27}$$

Proof To prove this lemma follow an argument used in Section 5.7 of [15], modified slightly to deal with the inhomogeneous boundary conditions.

Consider the problem: find $Q_h u_0 \in \mathcal{V}_{h,g}$ satisfying:

$$\left(F'(u_0)Q_h u_0, v_h \right) = \left(F'(u_0)u_0, v_h \right), \quad v_h \in \mathcal{V}_{h,0}. \tag{4.2.28}$$

As remarked above there exists a unique solution $Q_h u_0$ to (4.2.28).

Pick a suitable positive constant, $K \in \mathbb{R}$ ($K > -\min_{x \in \Omega} f'(u_0(x))$, for example), which ensures that there exists a positive constant β such that:

$$\left(F'(u_0)v, v \right) + K(v, v) \geq \beta \| v \|_1^2, \quad v \in \mathcal{V}. \tag{4.2.29}$$

We note that for any $v_h \in \mathcal{V}_{h,0}$:

$$\left(F'(u_0)[Q_h u_0 - u_0], v_h \right) = 0. \tag{4.2.30}$$

Using this and (4.2.29) it follows that for any $v_h \in \mathcal{V}_{h,g}$:

$$\begin{aligned}
\beta \parallel Q_h u_0 - u_0 \parallel_1^2 &\leq \left(\mathrm{F}'(u_0)\left[Q_h u_0 - u_0\right], \left[Q_h u_0 - u_0\right]\right) + \\
&\quad K\left(\left[Q_h u_0 - u_0\right], \left[Q_h u_0 - u_0\right]\right) \\
&= \left(\mathrm{F}'(u_0)\left[Q_h u_0 - u_0\right], \left[v_h - u_0\right]\right) + \\
&\quad K \parallel Q_h u_0 - u_0 \parallel_0^2 \\
&\leq C \parallel Q_h u_0 - u_0 \parallel_1 \parallel u_0 - v_h \parallel_1 + \\
&\quad K \parallel Q_h u_0 - u_0 \parallel_0^2 . \qquad (4.2.31)
\end{aligned}$$

Throughout this argument C is a constant which may depend on u_0.

Now we bound $\parallel Q_h u_0 - u_0 \parallel_0$ using duality arguments. Let $w \in \mathcal{V}_0$ be the solution (which exists by assumption **(A4)**) to the adjoint problem:

$$\left(\mathrm{F}'(u_0) w, v\right) = \left(\left[Q_h u_0 - u_0\right], v\right), \quad v \in \mathcal{V}_0. \qquad (4.2.32)$$

Using the fact that $(Q_h u_0 - \Pi_h g_e) - (u_0 - g_e) \in \mathcal{V}_0$, the self adjointness of $\mathrm{F}'(u_0)$ and (4.2.32), it follows that:

$$\begin{aligned}
\|Q_h u_0 - u_0\|_0^2 &= \left(\left[Q_h u_0 - u_0\right], \left[Q_h u_0 - u_0\right]\right) \\
&= \left(\left[Q_h u_0 - u_0\right], \left[(Q_h u_0 - \Pi_h g_e) - (u_0 - g_e)\right]\right) + \\
&\quad \left(\left[Q_h u_0 - u_0\right], \left[\Pi_h g_e - g_e\right]\right) \\
&= \left(\mathrm{F}'(u_0)\left[(Q_h u_0 - \Pi_h g_e) - (u_0 - g_e)\right], w\right) + \\
&\quad \left(\left[Q_h u_0 - u_0\right], \left[\Pi_h g_e - g_e\right]\right). \qquad (4.2.33)
\end{aligned}$$

Now, from (4.2.30) we deduce that for all $w_h \in \mathcal{V}_{h,0}$

$$\begin{aligned}
0 &= \left(\mathrm{F}'(u_0)\left[Q_h u_0 - u_0\right], w_h\right) \\
&= \left(\mathrm{F}'(u_0)\left[(Q_h u_0 - \Pi_h g_e) - (u_0 - g_e)\right], w_h\right) \\
&\quad + \left(\mathrm{F}'(u_0)\left[\Pi_h g_e - g_e\right], w_h\right).
\end{aligned}$$

which taken together with (4.2.33) implies that for all $w_h \in \mathcal{V}_{h,0}$:

$$\|Q_h u_0 - u_0\|_0^2 = \left(\mathrm{F}'(u_0)\left[(Q_h u_0 - \Pi_h g_e) - (u_0 - g_e)\right], \left[w - w_h\right]\right) +$$

$$
\begin{aligned}
& ([Q_h u_0 - u_0], [\Pi_h g_e - g_e]) \; + \; \left(\mathrm{F}'(u_0)\,[g_e - \Pi_h g_e], w_h\right) \\
\leq \; & C \parallel (Q_h u_0 - \Pi_h g_e) - (u_0 - g_e) \parallel_1 \parallel w - w_h \parallel_1 \; + \\
& \|Q_h u_0 - u_0\|_0 \|\Pi_h g_e - g_e\|_0 \; + \; C\|\Pi_h g_e - g_e\|_1 \|w_h\|_1 . \quad (4.2.34)
\end{aligned}
$$

Picking w_h to be the 'quasi-interpolant' of w defined in [61], we can then use [61, Corollary 4.1] to deduce that $\|w_h\|_1 \leq \|w\|_1$ and inequality (4.2.15) to show that:

$$
\begin{aligned}
\|Q_h u_0 - u_0\|_0^2 \; \leq \; & C h^\alpha \parallel (Q_h u_0 - \Pi_h g_e) - (u_0 - g_e) \parallel_1 \parallel w \parallel_{1+\alpha} \; + \\
& \parallel Q_h u_0 - u_0 \parallel_0 \parallel \Pi_h g_e - g_e \parallel_0 + C\|\Pi_h g_e - g_e\|_1 \|w\|_1 . \quad (4.2.35)
\end{aligned}
$$

From Remark 4.1.4 and (4.2.32) it follows that there exists a constant C such that:

$$
\parallel w \parallel_1 \leq \parallel w \parallel_{1+\alpha} \leq C \parallel Q_h u_0 - u_0 \parallel_0 ,
$$

which, taken with (4.2.35), shows that

$$
\begin{aligned}
\parallel Q_h u_0 - u_0 \parallel_0 \; & \leq \; C \{ h^\alpha \parallel (Q_h u_0 - \Pi_h g_e) - (u_0 - g_e) \parallel_1 + \|\Pi_h g_e - g_e\|_1 \} \\
& \leq \; C \{ h^\alpha \parallel Q_h u_0 - u_0 \parallel_1 + \|\Pi_h g_e - g_e\|_1 \} .
\end{aligned}
$$

Since $g_e \in \mathrm{H}^2$ it follows that

$$
\parallel Q_h u_0 - u_0 \parallel_0 \leq C \{ h^\alpha \parallel Q_h u_0 - u_0 \parallel_1 + h\|g_e\|_2 \} . \quad (4.2.36)
$$

Hence (4.2.31) and (4.2.36) imply that for all $v_h \in \mathcal{V}_{h,g}$:

$$
\beta \|Q_h u_0 - u_0\|_1^2 \leq C \|Q_h u_0 - u_0\|_1 \{ \parallel u_0 - v_h \parallel_1 + h^\alpha \parallel Q_h u_0 - u_0 \parallel_1 + h\|g_e\|_2 \} ,
$$

which implies that if h is chosen small enough:

$$
\parallel Q_h u_0 - u_0 \parallel_1 \leq C \{ \parallel u_0 - v_h \parallel_1 \; + h\|g_e\|_2 \} . \quad (4.2.37)
$$

Then with $v_h = \Pi_h u_0$ we may use (4.2.13) to deduce that

$$
\parallel Q_h u_0 - u_0 \parallel_1 \leq C h \left\{ \parallel u_0 \parallel_{\mathrm{H}^2_{\mathrm{u}}} \; + \|g_e\|_2 \right\}
$$

as required. □

We are now in a position to prove the following prerequisite for Brouwer's Fixed Point Theorem:

Lemma 4.2.10 *For all h sufficiently small, Φ_h maps $\mathcal{B}_h$ into $\mathcal{B}_h$ and is continuous.*

Proof For $v_h \in \mathcal{B}_h$, consider $\Phi_h(v_h)$. By (4.2.24) and (4.2.25) we can deduce that:

$$\mathrm{F}'(u_0)\left(\Phi_h(v_h) + \Pi_h g_e - Q_h u_0\right) = R(u_0, v_h + \Pi_h g_e) \quad \in (\mathcal{V}_{h,0})' .$$

Lemma 4.2.3(ii) implies that $\Phi_h(v_h)$ exists.

Thus using Lemma 4.2.3(ii) we have

$$\| \Phi_h(v_h) + \Pi_h g_e - Q_h u_0 \|_1 \leq C \| R(u_0, v_h + \Pi_h g_e) \|_{(\mathcal{V}_{h,0})'} . \tag{4.2.38}$$

Now (4.2.23) yields

$$\| R(u_0, v_h + \Pi_h g_e) \|_{(\mathcal{V}_{h,0})'} \leq \| f(u_0) + f'(u_0)\left[v_h + \Pi_h g_e - u_0\right] - f(v_h + \Pi_h g_e) \|_0 \tag{4.2.39}$$

which we shall bound in terms of h and $\|u_0\|_{\mathrm{H}^2_W} + \|g_e\|_2$.

In particular we shall show below that

$$\|R(u_0, v_h + \Pi_h g_e)\|_{(\mathcal{V}_{h,0})'} \leq Ch^2 \left\{ \|u_0\|_{\mathrm{H}^2_W} + \|g_e\|_2 \right\}^2 ,$$

which taken together with (4.2.38) shows that, for h sufficiently small

$$\| \Phi_h(v_h) + \Pi_h g_e - Q_h u_0 \|_1 \leq h \left\{ \|u_0\|_{\mathrm{H}^2_W} + \|g_e\|_2 \right\} .$$

This proves that Φ_h maps $\mathcal{B}_h$ to $\mathcal{B}_h$.

To obtain the bound on $R(u_0, v_h + \Pi_h g_e)$ use Taylor's Theorem on the right hand side of (4.2.39):

$$f(u_0) + f'(u_0)\left[v_h + \Pi_h g_e - u_0\right] - f(v_h + \Pi_h g_e) = -\frac{1}{2}\left[v_h + \Pi_h g_e - u_0\right]^2 f''(\zeta_h). \tag{4.2.40}$$

where $\zeta_h(\boldsymbol{x})$ lies between $u_0(\boldsymbol{x})$ and $v_h(\boldsymbol{x}) + \Pi_h g_e(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$. We shall show that $\| \zeta_h \|_\infty$ is bounded independently of h. This will be true if both $\|v_h\|_\infty$ and $\|\Pi_h g_e\|_\infty$

can be bounded independently of h, as u_0 is certainly independent of h. Since Π_h is the standard interpolant onto piecewise linear functions, $\|\Pi_h g_e\|_\infty \leq \|g_e\|_\infty$. Therefore it remains to bound $v_h \in \mathcal{B}_h$. Again recalling that:

$$\|\Pi_h u_0\|_\infty \leq \|u_0\|_\infty \leq C\|u_0\|_{1+\alpha}, \tag{4.2.41}$$

it follows from the discrete Sobolev inequality of [75, Lemma 2.1] that

$$\begin{aligned} \| v_h \|_\infty \leq & \| v_h + \Pi_h g_e - \Pi_h u_0 \|_\infty + \| \Pi_h u_0 - \Pi_h g_e \|_\infty \\ \leq & \quad C \left(\log \left(\frac{1}{\underline{h}} \right) \right)^{\frac{1}{2}} \| v_h + \Pi_h g_e - \Pi_h u_0 \|_1 + C \| u_0 \|_{1+\alpha} + \|g_e\|_\infty. \end{aligned} \tag{4.2.42}$$

Moreover, for any $v_h \in \mathcal{B}_h$,

$$\begin{aligned} \| v_h + \Pi_h g_e - \Pi_h u_0 \|_1 & \leq \| v_h + \Pi_h g_e - Q_h u_0 \|_1 + \| Q_h u_0 - \Pi_h u_0 \|_1 \\ & \leq h \left\{ \|u_0\|_{\mathrm{H}^2_W} + \|g_e\|_2 \right\} + \| Q_h u_0 - \Pi_h u_0 \|_1 . \end{aligned} \tag{4.2.43}$$

By Definition 4.2.7 we have the following equality in $(\mathcal{V}_{h,0})'$:

$$\mathrm{F}'(u_0)\left[Q_h u_0 - \Pi_h u_0\right] = \mathrm{F}'(u_0)\left[u_0 - \Pi_h u_0\right].$$

Thus by Lemma 4.2.3(ii) and (4.2.12), we have:

$$\begin{aligned} \| Q_h u_0 - \Pi_h u_0 \|_1 & \leq C \| \mathrm{F}'(u_0)\left[Q_h u_0 - \Pi_h u_0\right] \|_{(\mathcal{V}_{h,0})'} \\ & = C \| \mathrm{F}'(u_0)\left[u_0 - \Pi_h u_0\right] \|_{(\mathcal{V}_{h,0})'} \\ & \leq C \| u_0 - \Pi_h u_0 \|_1 \\ & \leq Ch \| u_0 \|_{\mathrm{H}^2_W} . \end{aligned} \tag{4.2.44}$$

Now returning to (4.2.42) and using (4.2.43), (4.2.44) and **(M4)**, we obtain, for $v_h \in \mathcal{B}_h$,

$$\begin{aligned} \|v_h\|_\infty & \leq Ch \left(\log \left(\frac{1}{\underline{h}} \right) \right)^{\frac{1}{2}} \left[\|u_0\|_{\mathrm{H}^2_W} + \|g_e\|_2 \right] + C\|u_0\|_{1+\alpha} + \|g_e\|_\infty \\ & \leq C \left[\|u_0\|_{\mathrm{H}^2_W} + \|u_0\|_{1+\alpha} + \|g_e\|_2 + \|g_e\|_\infty \right], \quad \text{as } h \to 0. \end{aligned}$$

Hence, for all $v_h \in \mathcal{B}_h$, $\|v_h\|_\infty \leq C$ as $h \to 0$.

Now returning to (4.2.40), we observe that ζ_h is bounded independently of h in the L_∞-norm as $h \to 0$. Moreover, since $v_h \in \mathcal{B}_h$, it follows from (4.2.26) and Lemma 4.2.9 that

$$\begin{aligned} \|(v_h + \Pi_h g_e) - u_0\|_1 &\leq \|u_0 - Q_h u_0\|_1 + \|v_h + \Pi_h g_e - Q_h u_0\|_1 \\ &\leq Ch\left\{\|u_0\|_{\mathrm{H}^2_W} + \|g_e\|_2\right\}, \end{aligned}$$

which taken together with (4.2.39) and (4.2.40) implies that

$$\| R(u_0, v_h + \Pi_h g_e) \|_{(\mathcal{V}_{h,0})'} \leq Ch^2 \left\{\| u_0 \|_{\mathrm{H}^2_W} + \|g_e\|_2\right\}^2, \text{ as } h \to 0.$$

This establishes the desired bound on $R(u_0, v_h + \Pi_h g_e)$.

Therefore for h sufficiently small, (4.2.38) tells us that

$$\begin{aligned} \|\Phi_h(v_h) + \Pi_h g_e - Q_h u_0\|_1 &\leq (Ch)\left(h\left\{\|u_0\|_{\mathrm{H}^2_W} + \|g_e\|_2\right\}\right) \\ &\leq h\left\{\|u_0\|_{\mathrm{H}^2_W} + \|g_e\|_2\right\}, \end{aligned}$$

which shows that $\Phi_h : \mathcal{B}_h \to \mathcal{B}_h$, as required.

To complete the proof it remains to show that Φ_h is continuous, to do this assume that $v_h, w_h \in \mathcal{B}_h$. Then

$$\mathrm{F}'(u_0)\left[\Phi_h(v_h) - \Phi_h(w_h)\right] = R(u_0, v_h + \Pi_h g_e) - R(u_0, w_h + \Pi_h g_e).$$

By Lemma 4.2.3(ii):

$$\| \Phi_h(v_h) - \Phi_h(w_h) \|_1 \leq C \| R(u_0, v_h + \Pi_h g_e) - R(u_0, w_h + \Pi_h g_e) \|_{(\mathcal{V}_{h,0})'},$$

and since $f \in \mathcal{C}^2$, a simple calculation shows that the right-hand side approaches 0 as $v_h \to w_h$ in $\|\cdot\|_1$. This proves the continuity of Φ_h, completing the proof. □

We can now conclude that there exists a (locally) unique solution to the finite element problem (4.2.16). This is the subject of the next result.

Theorem 4.2.11 *For h sufficiently small, (4.2.16) has a locally unique finite element*

solution u_h *with*

$$\| u_0 - u_h \|_1 \leq Ch \left\{ \| u_0 \|_{H^2_W} + \| g_e \|_2 \right\}, \tag{4.2.45}$$

$$\| u_0 - u_h \|_\infty \to 0, \quad as\ h \to 0. \tag{4.2.46}$$

Proof By Lemma 4.2.10 and Theorem 4.2.6 Φ_h has a fixed point v_h and by the remarks following Definition 4.2.5 $u_h := v_h + \Pi_h g_e$ satisfies (4.2.16). This gives the existence of a solution to the finite element system (4.2.16).

Since $u_h - \Pi_h g_e \in \mathcal{B}_h$, the required estimate comes from Lemma 4.2.9 and (4.2.26):

$$\begin{aligned} \| u_0 - u_h \|_1 &\leq \|u_0 - Q_h u_0\|_1 + \|u_h - Q_h u_0\|_1 \\ &= \|u_0 - Q_h u_0\|_1 + \| [u_h - \Pi_h g_e] + \Pi_h g_e - Q_h u_0\|_1 \\ &\leq Ch \left\{ \| u_0 \|_{H^2_W} + \| g_e \|_2 \right\}. \end{aligned}$$

(4.2.46) follows from (4.2.45) by assumptions **(M3)** and **(M4)** and the discrete Sobolev inequality of [75].

Finally, we use the third part of Lemma 4.2.3 to prove the local uniqueness of the finite element solution.

Suppose $u_{h,1}$ and $u_{h,2}$ are two solutions of (4.2.16) and consider $u_{h,1} - \Pi_h g_e, u_{h,2} - \Pi_h g_e \in \mathcal{B}_h$. Then, for all $v_h \in \mathcal{V}_{h,0}$:

$$\begin{aligned} 0 &= (\nabla u_{h,1}, \nabla v_h) + (f(u_{h,1}), v_h) - (\nabla u_{h,2}, \nabla v_h) - (f(u_{h,2}), v_h) \\ &= (\nabla [u_{h,1} - u_{h,2}], \nabla v_h) + ([f(u_{h,1}) - f(u_{h,2})], v_h). \end{aligned}$$

By the mean value theorem (Section 3.2 of [54]), there exists a $t \in (0,1)$ such that

$$f(u_{h,1}) - f(u_{h,2}) = f'(u_{h,2} + t(u_{h,1} - u_{h,2}))(u_{h,1} - u_{h,2})$$

and since $u_{h,1} - \Pi_h g_e, u_{h,2} - \Pi_h g_e \in \mathcal{B}_h$, it follows that $u_{h,2} + t(u_{h,1} - u_{h,2}) - \Pi_h g_e \in \mathcal{B}_h$. Defining $u_{h,0} := u_{h,2} + t(u_{h,1} - u_{h,2})$, then for all $v_h \in \mathcal{V}_{h,0}$:

$$\begin{aligned} 0 &= (\nabla [u_{h,1} - u_{h,2}], \nabla v_h) + ([f(u_{h,1}) - f(u_{h,2})], v_h) \\ &= (\nabla [u_{h,1} - u_{h,2}], \nabla v_h) + (f'(u_{h,0})[u_{h,1} - u_{h,2}], v_h) \end{aligned}$$

$$= (\mathrm{F}'(u_{h,0})[u_{h,1} - u_{h,2}], v_h) \tag{4.2.47}$$

As $u_{h,0} - \Pi_h g_e \in \mathcal{B}_h$ it follows that

$$\|u_0 - u_{h,0}\|_1 \leq Ch\left\{\|u_0\|_{\mathrm{H}^2_W} + \| g_e \|_2\right\}$$

and using **(M4)** and the discrete Sobolev inequality that $\|u_0 - u_{h,0}\|_\infty \to 0$ as $h \to 0$. Hence we may choose h small enough such that

$$\| u_0 - u_{h,0} \|_\infty \leq \delta$$

where δ is as given in Lemma 4.2.3.

Thus from the third part of Lemma 4.2.3:

$$\begin{aligned} \| u_{h,1} - u_{h,2} \|_1 &\leq C(\delta) \| \mathrm{F}'(u_{h,0})[u_{h,1} - u_{h,2}] \|_{(\mathcal{V}_{h,0})'} \\ &= 0 \end{aligned}$$

by (4.2.47). □

It follows from this theorem that the locally unique solution of (4.2.16) is bounded independently of h in the L_∞-norm as $h \to 0$.

4.3 A Multilevel Adaptive Scheme

A simple strategy for solving (4.1.1)-(4.1.3) to a required tolerance, TOL, would be to use a refinement strategy. This could involve solving (4.2.17) for some sequence of triangulations, $\{\mathcal{T}_h^k\}$, with decreasing mesh size h, $h^0 > h^1 > h^2 > \ldots$. [h^k denotes the maximum diameter of the triangles in the triangulation $\mathcal{T}_h^k$ and **not h to the power k**]. We denote the corresponding finite element space by $\mathcal{V}_h^k$ and define the standard finite element solution to (4.2.17) in this space to be $u_h^k, k \geq 0$. We might accept a finite element solution, u_h^K, as good enough if, for example:

$$\| u_h^K - u_h^{K-1} \|_1 \leq \mathrm{TOL} \tag{4.3.48}$$

We know from (4.2.45) that for K sufficiently large (4.3.48) will eventually be satisfied.

If $f(u)$ is linear in u, then the cost of finding a finite element solution satisfying (4.3.48) is the cost of solving a sequence of linear problems for each k.

However if $f(u)$ is nonlinear we have to solve a sequence of nonlinear problems, each of which must be solved by some inner iteration, by Newton's Method for example. This is much more expensive than the cost of solving a sequence of linear problems. We aim to introduce a method for solving our semilinear problem with approximately the same cost as solving a problem with linear f.

We describe a method for computing an approximate sequence:

$$\hat{u}_h^k \cong u_h^k, \quad k = 0, 1, 2, \ldots,$$

with starting value:

$$\hat{u}_h^0 := u_h^0,$$

such that $\hat{u}_h^k$ is much cheaper to compute than u_h^k, but an optimal error estimate remains true:

$$\| u_0 - \hat{u}_h^k \|_1 = O(h^k). \tag{4.3.49}$$

Thus we will still be able to find a solution, $\hat{u}_h^K$, such that (4.3.48) is satisfied, for sufficiently large K. It turns out that the cost of computing this sequence is the cost of solving one nonlinear problem on the coarsest mesh (i.e. $k = 0$), plus the cost of one linear problem for each $k = 1, 2, \ldots, K$. The total work is thus not much more than in the linear case. We show that if the coarsest mesh diameter, h^0, is sufficiently fine, then (4.3.49) is satisfied for the sequence of approximations $\{\hat{u}_h^k\}$ which we shall define below. We call our method a *multilevel defect correction* scheme.

4.3.1 Related Work

The methods described in this chapter are strongly related to the work of Xu in [72], [73] and Axelsson in [6]. In [72] a two mesh method is proposed where the solution to the finite element discretisation of a semilinear problem is found on a coarse mesh and then a correction is calculated using one step of a Newton iteration (a linear solve) on the finer mesh. The original finite element solution is then updated before a final correction is found on the coarse mesh (another linear solve). The algorithm proposed in [6] is essentially the same, but does not include the final coarse mesh solve. In [73,

Section 5.4] the method of [72] is extended to the situation with more than one fine mesh in much the same way as is proposed here.

In [72], [73] and [6] very good convergence rates are obtained for the methods, providing the maximum mesh diameters of the meshes decrease sufficiently rapidly. In this chapter we consider much weaker links between the meshes. Xu considers the semilinear problem set on a convex domain and assumes the existence of the finite element solution. Axelsson assumes that the meshes are quasi uniform and the derivative of the function f with respect to u is positive (which guarantees the existence and uniqueness of the finite element solution, see for example [42]). The results presented here are more general in that they cover problems with irregular solutions (i.e. the solutions have singularities induced by corners/mixed boundary conditions and f is allowed to be non-monotone).

Other related work can be found in [3], [4] and the references therein. These propose a multilevel method (the "Discrete Defect Correction Method") which tries to minimise the amount of work needed on the intermediate meshes. They try to exploit the properties of the Mesh Independence Principle (MIP, see for example [3]) to limit most of the work to the initial and final meshes. The MIP is basically the idea that for any mesh with small enough mesh diameter Newton's method will take a fixed number of steps to converge. There are various conditions to check if the MIP holds and these can be found in [3]. In [3] and [4] the following algorithm is proposed:

1. Choose a coarse mesh.

2. Solve the nonlinear problem on the current mesh using Newton's method.

3. Refine the mesh and start resolving on the new mesh. If the MIP holds go to step 4, if not refine the mesh and return to 2.

4. Choose a sequence of finer meshes. Perform one step of Newton's method on each of the finer meshes to update the solution using the previous solution interpolated onto the new mesh as an initial guess. Finally on the finest mesh solve the problem to full tolerance.

There is no proof that the algorithm is well defined or that it will converge to the required solution. If the choice of initial mesh is suitable (i.e. it satisfies the MIP), then

the method is close to the Defect Correction method considered here. Otherwise the connection is not so obvious.

4.3.2 The Defect Correction Algorithm

For each k, let F_h^k denote the nonlinear map $\mathrm{F}_h^k : \mathcal{V}_{h,g}^k \to \left(\mathcal{V}_{h,0}^k\right)'$ defined by replacing $\mathcal{V}_h$ by $\mathcal{V}_h^k$ in (4.2.17). Let $\left(\mathrm{F}_h^k\right)' : \mathcal{V}_{h,g}^k \to \mathrm{L}\left(\mathcal{V}_{h,0}^k, \left(\mathcal{V}_{h,0}^k\right)'\right)$ denote the Jacobian of F_h^k, where $L(X,Y)$ denotes the space of linear, continuous maps, $A : X \to Y$.

Our Defect Correction Algorithm is:

1. Set $\hat{u}_h^0 = u_h^0$, the exact solution in $\mathcal{V}_{h,g}^0$ of the nonlinear finite element problem $\mathrm{F}_h^0(u_h^0) = 0$ in $\left(\mathcal{V}_{h,0}^0\right)'$.

2. For $k = 0, 1, 2, \ldots$, iterate the two steps:

   - Solve for $\hat{e}_h^{k+1} \in \mathcal{V}_{h,0}^{k+1}$:

$$\left(\mathrm{F}_h^{k+1}\right)'(\hat{u}_h^k)\ \hat{e}_h^{k+1} = -\mathrm{F}_h^{k+1}(\hat{u}_h^k). \tag{4.3.50}$$

   - Update $\hat{u}_h^k$:

$$\hat{u}_h^{k+1} = \hat{u}_h^k + \hat{e}_h^{k+1}. \tag{4.3.51}$$

Step 2 can be considered to be a sequence of single steps of Newton's Method on successively finer meshes. Note that since $\mathcal{T}_h^{k+1}$ is a refinement of $\mathcal{T}_h^k$, $\mathcal{V}_h^k \subset \mathcal{V}_h^{k+1}$ and (4.3.51) defines an update in $\mathcal{V}_h^{k+1}$.

Since $\mathcal{V}_h^k \subset \mathcal{C}(\Omega)$, $\left(\mathrm{F}_h^{k+1}\right)'(\hat{u}_h^k)$ and $\mathrm{F}_h^{k+1}(\hat{u}_h^k)$ are well defined for all k, all that remains is to show is that (4.3.50) is solvable in order for the algorithm to be well-defined. This is shown in Lemma 4.3.2. As a preparation for the proof we need the following result:

Lemma 4.3.1 *Suppose $u \in \mathcal{V}_{h,g} \cap L_\infty$ with $\| u_0 - u \|_\infty \leq \delta$, where δ is as given in Lemma 4.2.3, then for all $b \in L_2$, there exists a unique solution $v_h^k \in \mathcal{V}_{h,0}^k$ such that:*

$$\left(\left(F_h^k\right)'(u)\ v_h^k, w_h^k\right) = \left(b, w_h^k\right)$$

for all $w_h^k \in \mathcal{V}_{h,0}^k$.

Proof We note that finding v_h^k is equivalent to solving a square linear system, thus uniqueness implies existence. To show uniqueness, suppose

$$\left(\left(\mathrm{F}_h^k\right)'(u)\, v_h^k, w_h^k\right) = 0, \quad w_h^k \in \mathcal{V}_{h,0}^k$$

then Lemma 4.2.3(ii) implies that:

$$\begin{aligned} \| v_h^k \|_1 &\leq C \| \left(\mathrm{F}_h^k\right)'(u)\, v_h^k \|_{(\mathcal{V}_{h,0}^k)'} \\ &= 0. \end{aligned}$$

Thus $v_h^k = 0$, completing the proof. □

4.3.3 Convergence of the Defect Correction Method

It is now possible to prove that the algorithm is well defined and that the defect correction method converges. It will be shown that, providing the meshes are refined sufficiently cautiously, the sequence of defect correction solutions satisfy an error estimate of the following form:

$$\|u_0 - \hat{u}_h^k\|_1 \leq C_1 \left(1 + C_2 \left(h^k\right)^\epsilon\right) h^k \left\{\|u_0\|_{\mathrm{H}_W^2} + \| g_e \|_2\right\}, \quad k = 0, 1, \ldots, \tag{4.3.52}$$

where C_1 is the constant appearing in the error estimate for the exact finite element solution u_h^k, as in (4.2.45). ϵ and C_2 are fixed constants, independent of k, the identity of which will be described below.

The next lemma shows that the algorithm is well defined and also contains a key step in the proof that the defect correction method converges:

Lemma 4.3.2 *Suppose $\hat{u}_h^k \in \mathcal{V}_h^k$ satisfies*

$$\| u_0 - \hat{u}_h^k \|_\infty \leq \delta, \tag{4.3.53}$$

where δ is as given in Lemma 4.2.3. Then $\hat{u}_h^{k+1} \in \mathcal{V}_{h,g}^{k+1}$ is well defined. Further, for all $p \in (2, \infty)$, there exists a constant C_p such that

$$\| \hat{u}_h^{k+1} - u_h^{k+1} \|_1 \leq C_p \| \hat{u}_h^k - u_h^{k+1} \|_{\mathrm{L}_p}^2. \tag{4.3.54}$$

Proof If (4.3.53) is satisfied, then it follows from Lemma 4.3.1 that $\hat{u}_h^{k+1} \in \mathcal{V}_{h,g}^{k+1}$, given by (4.3.50) and (4.3.51), is well-defined.

Moreover, since $\hat{e}_h^{k+1} = \hat{u}_h^{k+1} - \hat{u}_h^k$, (4.3.50) may be rearranged to give:

$$\left(\nabla \hat{u}_h^{k+1}, \nabla w_h^{k+1}\right) + \left(f(\hat{u}_h^k) + f'(\hat{u}_h^k)\left[\hat{u}_h^{k+1} - \hat{u}_h^k\right], w_h^{k+1}\right) = 0,$$

for all $w_h^{k+1} \in \mathcal{V}_{h,0}^{k+1}$. However, recalling that $u_h^{k+1} \in \mathcal{V}_{h,g}^{k+1}$ is the exact solution of (4.2.17) with $h = h^{k+1}$, we have:

$$\left(\nabla u_h^{k+1}, \nabla w_h^{k+1}\right) + \left(f(u_h^{k+1}), w_h^{k+1}\right) = 0.$$

Hence, subtracting and rearranging these two expressions shows that:

$$\left(\mathrm{F}_h^{k+1}\right)'(\hat{u}_h^k)\left[\hat{u}_h^{k+1} - u_h^{k+1}\right] = \left(f(u_h^{k+1}) - f(\hat{u}_h^k) - f'(\hat{u}_h^k)\left[u_h^{k+1} - \hat{u}_h^k\right]\right), \text{ in } \left(\mathcal{V}_{h,0}^{k+1}\right)'.$$

The right hand side has to be understood as an element of $\left(\mathcal{V}_{h,0}^{k+1}\right)'$ in the right way, i.e. as the linear functional

$$w_h^{k+1} \to \left(f(u_h^{k+1}) - f(\hat{u}_h^k) - f'(\hat{u}_h^k)\left[u_h^{k+1} - \hat{u}_h^k\right], w_h^{k+1}\right).$$

Here we have the usual L_2 inner product, not the H^1 inner product.

Then Lemma 4.2.3(ii) implies that:

$$\begin{aligned} \| \hat{u}_h^{k+1} - u_h^{k+1} \|_1 &\leq C \| \left(\mathrm{F}_h^{k+1}\right)'(\hat{u}_h^k)\left[\hat{u}_h^{k+1} - u_h^{k+1}\right] \|_{\left(\mathcal{V}_{h,0}^{k+1}\right)'} \\ &= C \| f(u_h^{k+1}) - f(\hat{u}_h^k) - f'(\hat{u}_h^k)\left[u_h^{k+1} - \hat{u}_h^k\right] \|_{\left(\mathcal{V}_{h,0}^{k+1}\right)'} \end{aligned} \tag{4.3.55}$$

Now observe that if b is any function in L_∞, then by definition

$$\| b \|_{\left(\mathcal{V}_{h,0}^{k+1}\right)'} := \sup_{w_h \in \mathcal{V}_{h,0}^{k+1},\ w_h \neq 0} \frac{| (b, w_h) |}{\| w_h \|_1}$$

and by Hölder's inequality

$$| (b, w_h) | \leq \| b \|_{\mathrm{L}_{p/2}} \| w_h \|_{\mathrm{L}_q}.$$

where $\frac{2}{p}+\frac{1}{q}=1$ and $p\in(2,\infty)$, $q\in(1,\infty)$. Since the Sobolev Embedding Theorem tells us that $\|w_h\|_{\mathrm{L}_q}\leq C\|w_h\|_1$, it follows that

$$\begin{aligned}\| b \|_{(\mathcal{V}_{h,0}^{k+1})'} &\leq \| b \|_{\mathrm{L}_{p/2}} \\ &= \left\{\| |b|^{\frac{1}{2}} \|_{\mathrm{L}_p}\right\}^2 \end{aligned} \tag{4.3.56}$$

Then to estimate (4.3.55) use Taylor's Theorem to deduce that there exists a $\xi(\boldsymbol{x})$, lying between $u_h^{k+1}(\boldsymbol{x})$ and $\hat{u}_h^k(\boldsymbol{x})$ for all $\boldsymbol{x}\in\Omega$, such that

$$f(u_h^{k+1})-f(\hat{u}_h^k)-f'(\hat{u}_h^k)\left[u_h^{k+1}-\hat{u}_h^k\right]=f''(\xi)\left[u_h^{k+1}-\hat{u}_h^k\right]^2.$$

Since it has been assumed that $\|u_0-\hat{u}_h^k\|_\infty\leq\delta$ it follows that

$$\| \hat{u}_h^k \|_\infty \leq \delta + \| u_0 \|_\infty$$

and because u_h^{k+1} is also bounded independently of k [this follows from Theorem 4.2.11: note that h^k is the diameter of the kth mesh], it follows that ξ is similarly bounded. Thus there exists a constant C, independent of k, such that:

$$\| \left|f(u_h^{k+1})-f(\hat{u}_h^k)-f'(\hat{u}_h^k)\left[u_h^{k+1}-\hat{u}_h^k\right]\right|^{\frac{1}{2}} \|_{\mathrm{L}_p} \leq C \| u_h^{k+1}-\hat{u}_h^k \|_{\mathrm{L}_p}$$

Taking this together with (4.3.55) and (4.3.56) proves the required estimate. □

As indicated above, the defect correction method only performs well when the meshes are refined in a sufficiently careful way, that is to say there is a limit on the number of new mesh points which can be introduced at each refinement step. This is reflected in the following new mesh refinement assumption:

(M5) There exists positive constants $\gamma>0$ and $\epsilon\in(0,1)$, independent of k, such that for all k:

$$\left(h^k\right)^2\leq\gamma\left(h^0\right)^{1-\epsilon}\left(h^{k+1}\right)^{1+\epsilon}. \tag{4.3.57}$$

Remark 4.3.3 *This essentially requires $(h^k)^2\leq C(h^{k+1})^{1+\epsilon}$, but we choose the more specific form of the constant for convenience.*

With mesh condition **(M5)** it is possible to prove the following theorem:

Theorem 4.3.4 *If h^0 is sufficiently small and the sequence of meshes satisfy (4.3.57), then the defect correction solution sequence $\{\hat{u}_h^k : k = 1, 2, \ldots\}$ is well defined and satisfies*

$$\|u_0 - \hat{u}_h^k\|_1 \leq C_1 \left(1 + C_2 \left(h^k\right)^{\epsilon}\right) h^k \left\{\|u_0\|_{\mathrm{H}^2_W} + \| g_e \|_2\right\}, \quad k = 0, 1, \ldots. \tag{4.3.58}$$

C_2 is a constant given by:

$$C_2 = 2C_p C_1 \left[4\gamma + 1\right] \left(h^0\right)^{1-\epsilon} \left\{\|u_0\|_{\mathrm{H}^2_W} + \| g_e \|_2\right\}, \tag{4.3.59}$$

C_1 is the constant appearing in (4.2.45), ϵ and γ are the constants appearing in (4.3.57) and C_p is the constant in (4.3.54).

Proof The result is proved by induction.

Assume that the result is true for k. If it can be shown that

$$\|u_0 - \hat{u}_h^k\|_\infty \leq \delta \tag{4.3.60}$$

holds, then if follows from Lemma 4.3.2 that $\hat{u}_h^{k+1}$ is well defined and satisfies

$$\| \hat{u}_h^{k+1} - u_h^{k+1} \|_1 \leq C_p \| \hat{u}_h^k - u_h^{k+1} \|^2_{\mathrm{L}_p}. \tag{4.3.61}$$

for $p \in (2, \infty)$.

To prove (4.3.60), use the discrete Sobolev inequality of [75], the interpolation assumptions (4.2.12), (4.2.14) and the inductive hypothesis that (4.3.58) is true for $\hat{u}_h^k$, to show that

$$\begin{aligned}
\|u_0 - \hat{u}_h^k\|_\infty &\leq \|u_0 - \Pi_h^k u_0\|_\infty + \|\Pi_h^k u_0 - \hat{u}_h^k\|_\infty \\
&\leq C\left(h^k\right)^2 \|u_0\|_{\mathrm{C}^2_W} + C'\left(\log\left(\frac{1}{\underline{h}^k}\right)\right)^{\frac{1}{2}} \|\Pi_h^k u_0 - \hat{u}_h^k\|_1 \\
&\leq C\left(h^k\right)^2 \|u_0\|_{\mathrm{C}^2_W} + C'\left(\log\left(\frac{1}{\underline{h}^k}\right)\right)^{\frac{1}{2}} \left[Ch^k \|u_0\|_{\mathrm{H}^2_W} + \|u_0 - \hat{u}_h^k\|_1\right] \\
&\leq C\left(h^k\right)^2 \|u_0\|_{\mathrm{C}^2_W} + \\
&\quad C\left(\log\left(\frac{1}{\underline{h}^k}\right)\right)^{\frac{1}{2}} \left[Ch^k + C_1\left(1 + C_2\left(h^k\right)^{\epsilon}\right) h^k\right] \left\{\|u_0\|_{\mathrm{H}^2_W} + \| g_e \|_2\right\}
\end{aligned}$$

$$\leq \quad Ch^k \left(\log \left(\frac{1}{\underline{h}^k} \right) \right)^{\frac{1}{2}}. \tag{4.3.62}$$

Since it follows from **(M4)** that

$$h^k \left(\log \left(\frac{1}{\underline{h}^k} \right) \right)^{\frac{1}{2}} \leq h^0 \left(\log \left(\frac{1}{\underline{h}^0} \right) \right)^{\frac{1}{2}},$$

(4.3.60) holds if h^0 is taken sufficiently small such that

$$Ch^0 \left(\log \left(\frac{1}{\underline{h}^0} \right) \right)^{\frac{1}{2}} \leq \delta, \tag{4.3.63}$$

where C is the constant appearing in (4.3.62).

Therefore, for h^0 sufficiently small,

$$\| u_0 - \hat{u}_h^k \|_\infty \leq \delta$$

and $\hat{u}_h^{k+1}$ is well defined and satisfies (4.3.61).

Now, to show that (4.3.58) is true for $k = k + 1$, use the triangle inequality and the error estimate for the exact finite element solution (4.2.45) on the $k + 1$th mesh:

$$\begin{aligned} \| u_0 - \hat{u}_h^{k+1} \|_1 &\leq \| u_0 - u_h^{k+1} \|_1 + \| u_h^{k+1} - \hat{u}_h^{k+1} \|_1 \\ &\leq C_1 h^{k+1} \left\{ \| u_0 \|_{\mathrm{H}^2_W} + \| g_e \|_2 \right\} + \| u_h^{k+1} - \hat{u}_h^{k+1} \|_1. \end{aligned} \tag{4.3.64}$$

Comparing (4.3.64) with (4.3.58) we see that it remains to show that

$$\| u_h^{k+1} - \hat{u}_h^{k+1} \|_1 \leq C_1 C_2 \left(h^{k+1} \right)^{1+\epsilon} \left\{ \| u_0 \|_{\mathrm{H}^2_W} + \| g_e \|_2 \right\}. \tag{4.3.65}$$

With this in mind, consider the left hand side of (4.3.65) and apply mesh assumption **(M5)** and (4.3.61), for $p \in (2, \infty)$:

$$\begin{aligned} \| u_h^{k+1} - \hat{u}_h^{k+1} \|_1 &\leq C_p \| u_h^{k+1} - \hat{u}_h^k \|_{1,p}^2 \\ &\leq C_p \| u_h^{k+1} - \hat{u}_h^k \|_1^2 \\ &\leq 2C_p \left[\| u_h^{k+1} - u_0 \|_1^2 + \| u_0 - \hat{u}_h^k \|_1^2 \right] \\ &\leq 2C_p \left[C_1^2 \left(h^{k+1} \right)^2 + C_1^2 \left(h^k \right)^2 \left(1 + C_2 \left(h^k \right)^\epsilon \right)^2 \right] \left\{ \| u_0 \|_{\mathrm{H}^2_W} + \| g_e \|_2 \right\}^2 \end{aligned}$$

$$= 2C_pC_1^2\left[\left(h^{k+1}\right)^{1-\epsilon}+\gamma\left(h^0\right)^{1-\epsilon}\left(1+C_2\left(h^k\right)^{\epsilon}\right)^2\right]\left(h^{k+1}\right)^{1+\epsilon}\left\{\|u_0\|_{\mathrm{H}^2_W}+\|\,g_e\,\|_2\right\}^2.$$

Since $h^0 \geq h^k \geq h^{k+1}$, it follows that

$$\|u_h^{k+1}-\hat{u}_h^{k+1}\|_1 \leq 2C_pC_1^2\left[1+\gamma\left(1+C_2\left(h^0\right)^{\epsilon}\right)^2\right]\left(h^0\right)^{1-\epsilon}\left(h^{k+1}\right)^{1+\epsilon}\left\{\|u_0\|^2_{\mathrm{H}^2_W}+\|\,g_e\,\|_2\right\}. \tag{4.3.66}$$

Taking h^0 sufficiently small such that

$$2C_pC_1\left[4\gamma+1\right]\left\{\|u_0\|_{\mathrm{H}^2_W}+\|\,g_e\,\|_2\right\} \leq \left(h^0\right)^{-1}$$

it follows from the definition of C_2, (4.3.59), that:

$$C_2\left(h^0\right)^{\epsilon} = 2C_pC_1\left[4\gamma+1\right]\left\{\|u_0\|_{\mathrm{H}^2_W}+\|\,g_e\,\|_2\right\}\left(h^0\right)^1 \leq 1.$$

Therefore from (4.3.66) we have

$$\begin{aligned}\|u_h^{k+1}-\hat{u}_h^{k+1}\|_1 &\leq 2C_pC_1^2\left[1+4\gamma\right]\left(h^0\right)^{1-\epsilon}\left(h^{k+1}\right)^{1+\epsilon}\left\{\|u_0\|_{\mathrm{H}^2_W}+\|\,g_e\,\|_2\right\}^2\\ &= C_1C_2\left(h^{k+1}\right)^{1+\epsilon}\left\{\|u_0\|_{\mathrm{H}^2_W}+\|\,g_e\,\|_2\right\}\end{aligned} \tag{4.3.67}$$

proving (4.3.65) holds. Therefore combining (4.3.64) and (4.3.67), we obtain

$$\|u_0-\hat{u}_h^{k+1}\|_1 \leq C_1\left(1+C_2\left(h^{k+1}\right)^{\epsilon}\right)h^{k+1}\left\{\|u_0\|_{\mathrm{H}^2_W}+\|\,g_e\,\|_2\right\}$$

as required.

We conclude that if the result holds for k, then it holds for $k+1$. To complete the proof it remains to show that the result holds for $k=0$.

Since $\hat{u}_h^0 = u_h^0$, $\hat{u}_h^0$ is well defined by Theorem 4.2.11. From the discrete Sobolev inequality of [75], (4.2.45), (4.2.12) and (4.2.14):

$$\begin{aligned}\|u_0-\hat{u}_h^0\|_\infty &= \|u_0-u_h^0\|_\infty\\ &\leq \|u_0-\Pi_h^0u_0\|_\infty+\|\Pi_h^0u_0-u_h^0\|_\infty\\ &\leq C\left(h^0\right)^2\|u_0\|_{C^2_W}+C\left(\log\left(\frac{1}{h^0}\right)\right)^{\frac{1}{2}}\|\Pi_h^0u_0-u_h^0\|_1\\ &\leq C\left(h^0\right)^2\|u_0\|_{C^2_W}+C\left(\log\left(\frac{1}{h^0}\right)\right)^{\frac{1}{2}}h^0\left\{\|u_0\|_{\mathrm{H}^2_W}+\|\,g_e\,\|_2\right\}\end{aligned}$$

$$\leq \quad Ch^0 \left(\log \left(\frac{1}{\underline{h}^0} \right) \right)^{\frac{1}{2}}$$

which implies

$$\|u_0 - \hat{u}_h^0\|_\infty \leq \delta$$

by (4.3.63).

$\hat{u}_h^0$ obviously satisfies the error estimate (4.3.58) as

$$\begin{aligned} \|u_0 - \hat{u}_h^0\|_1 &= \|u_0 - u_h^0\|_1 \\ &\leq C_1 h^0 \left\{ \|u_0\|_{\mathrm{H}^2_W} + \| g_e \|_2 \right\}. \end{aligned}$$

This completes the inductive proof. □

The following corollary compares the error in the standard finite solution with the error in the defect correction solution. It is shown that the errors in the H^1-norm on the kth mesh are the same, up to order $\left(h^k\right)^{1+\epsilon}$.

Corollary 4.3.5 *If h^0 is sufficiently small and the sequence of meshes satisfy (4.3.57), then:*

$$\|u_0 - u_h^k\|_1 - O\left(\left(h^k\right)^{1+\epsilon}\right) \leq \|u_0 - \hat{u}_h^k\|_1 \leq \|u_0 - u_h^k\|_1 + O\left(\left(h^k\right)^{1+\epsilon}\right). \qquad (4.3.68)$$

Proof Using the triangle inequality

$$\|u_0 - u_h^k\|_1 - \|u_h^k - \hat{u}_h^k\|_1 \leq \|u_0 - \hat{u}_h^k\|_1 \leq \|u_0 - u_h^k\|_1 + \|u_h^k - \hat{u}_h^k\|_1. \qquad (4.3.69)$$

From (4.3.67) it follows that

$$\|u_h^k - \hat{u}_h^k\|_1 \leq C\left(h^k\right)^{1+\epsilon} \left\{ \|u_0\|_{\mathrm{H}^2_W} + \| g_e \|_2 \right\}$$

which combined with (4.3.69) implies the result. □

Chapter 5

A Posteriori Error Estimates for Semilinear Elliptic Equations

In this chapter we study *a posteriori* error estimates and adaptive methods for semilinear problems. An *a posteriori* error estimate is a computable bound on the error in an approximate solution to a problem. It usually involves the approximate solution which has already been found. As we have seen in the previous chapter semilinear equations often arise in semiconductor modelling. As an initial insight into why we study adaptive methods consider the semiconductor equations (1.3.12)-(1.3.14) under zero applied voltage and $\lambda = 0$. With such an assumption $v = w = 0$ and the system reduces to finding ψ such that $2\delta^2 \sinh\psi - d = 0$. This has exact solution $\psi = \sinh^{-1}(d/2\delta^2)$, which has the same jumps as d itself. For small λ these jumps become narrow interior layers (regions of fast variation in the solution). As the voltage is increased v and w are no longer identically equal to zero but also have narrow layers in the vicinity of the jumps in d. There is a large literature that describes the asymptotics of the solutions to the semiconductor equations as $\lambda \to 0$ and/or $\delta \to 0$ (see for example [56] or [51]). The aim of this chapter will be to introduce *a posteriori* error estimators and an adaptive method that will be capable of fully capturing these interior layers. The *a posteriori* error estimate and adaptive scheme will be combined with the defect correction method of Chapter 4 to form an adaptive defect correction method in Chapter 6.

Although quite a lot of work (see Section 1.4) has been done on the adaptive solution of the semiconductor problem, much of it is not rigorous or is based on *a priori* information. Here we give rigorous a posteriori error estimates and demonstrate that an

adaptive scheme based on the estimate is capable of capturing all the detail of a model semiconductor problem with known asymptotic solution.

Appearing in the *a posteriori* error estimates are constants independent of the mesh parameters and finite element solution. Rather than analysing these constants in detail and giving bounds on their size we instead give a method of numerically estimating them. We demonstrate that this estimation scheme works well and find that in some cases the effective values of these constants may be smaller than a purely analytic theory would predict.

Although a lot of work has been done on finding *a posteriori* error estimates for linear problems (see for example [7], [11], [32], [14], [31]), work on nonlinear examples has been considerably scarcer. *A posteriori* error estimators for general nonlinear problems are presented in [57], [70], [71] and [12]. [57] gives H^1 *a posteriori* error estimators and we will follow a similar procedure and extend the results to the L_2-norm. Verfürth produces H^1 *a posteriori* error estimates in [70] and $L_r, r \in (1, \infty)$ estimates in [71] for a very general class of problems on polygonal domains with Dirichlet boundary conditions. These estimates are very general, but are quite difficult to use due to the presence of approximation terms. Our *a posteriori* error estimates will be of an analogous form, but will avoid the use of these approximating terms. Similar results to Verfürth in the L_2-norm are obtained in [12].

5.1 The Semilinear Problem Considered

We consider semilinear problems of the following type:

$$-\Delta u(\boldsymbol{x}) + f(u(\boldsymbol{x}), \boldsymbol{x}) = 0, \qquad \text{in } \Omega, \tag{5.1.1}$$

$$u = g, \qquad \text{on } \partial\Omega_D. \tag{5.1.2}$$

$$\frac{\partial u}{\partial n} = 0, \qquad \text{on } \partial\Omega_N. \tag{5.1.3}$$

We assume that Ω is a bounded polygonal domain in $\mathbb{R}^2$ and that $\partial\Omega$ can be decomposed as the union of $\partial\Omega_D$ and $\partial\Omega_N$, where $\partial\Omega_D$ and $\partial\Omega_N$ are disjoint sets consisting of line segments of $\partial\Omega$. We assume that $\partial\Omega_D \neq \emptyset$.

We further assume that

(A1) $f : \mathbb{R} \times \Omega \to \mathbb{R}$ has the property that for all $\boldsymbol{x} \in \Omega$, $f(\cdot, \boldsymbol{x}) \in C^2(\mathbb{R})$ and if

$u \in C(\Omega)$ then the function $\boldsymbol{x} \to f(u(\boldsymbol{x}), \boldsymbol{x})$ is in $\mathrm{L}_\infty(\Omega)$.

(A2) $g \in \mathrm{H}^{\frac{3}{2}}(\partial\Omega_D)$.

We use C and C' to denote generic positive constants whose numerical values may change from line to line. For convenience introduce the shorthand notation $f(u) := f(u(\boldsymbol{x}), \boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, $f'(u)$ and $f''(u)$ will be taken to mean the derivatives of f with respect to u.

Since we are interested in finding an accurate finite element approximation to the weak solution of (5.1.1)-(5.1.3) we need to define a family of finite element triangulations $\mathcal{T}_h$, each consisting of triangles T_k. We make the following restrictions on the type of underlying triangulation allowed. Assume that $\mathcal{T}_h$ satisfies:

(M1) $\Omega = \bigcup_{\mathrm{T}_k \in \mathcal{T}_h} \mathrm{T}_k$.

(M2) $\mathrm{T}_1, \mathrm{T}_2 \in \mathcal{T}_h$, $\mathrm{T}_1 \neq \mathrm{T}_2$, are either disjoint or have a vertex in common, or a side in common.

(M3) The triangulations are non-degenerate over the whole family.

(M4) We assume for each triangle, T_k, in the triangulation there are at most K triangles having a non-empty intersection with this triangle and that K is independent of the maximum triangle diameter of the triangulation.

(M5) No edge of a triangle on the boundary of Ω has both Dirichlet and Neumann boundary conditions defined on it (so if τ is an edge of a triangle on the boundary then either $\tau \in \partial\Omega_D$ or $\tau \in \partial\Omega_N$).

Remark 5.1.1 *Assumption (M4) is equivalent to (M3), but we leave it in as it helps the clarity of the proofs of the a posteriori error estimates.*

For every triangle T_k in our triangulation, $\mathcal{T}_h$, we define $\mathcal{E}(\mathrm{T}_k)$ to be the set of edges of our triangle. Let

$$\mathcal{E}_h = \bigcup_{\mathrm{T}_k \in \mathcal{T}_h} \mathcal{E}(\mathrm{T}_k).$$

We split $\mathcal{E}_h$ into three different sets:

$$\mathcal{E}_D := \{\tau \in \mathcal{E}_h : \tau \subset \partial\Omega_D\}.$$

$$\begin{aligned}\mathcal{E}_N &:= \{\tau \in \mathcal{E}_h : \tau \subset \partial\Omega_N\},\\ \mathcal{E}_\Omega &:= \mathcal{E}_h \setminus \{\mathcal{E}_D \cup \mathcal{E}_N\},\end{aligned}$$

thus $\mathcal{E}_h = \mathcal{E}_\Omega \cup \mathcal{E}_D \cup \mathcal{E}_N$. We define the mesh parameters:

$$\begin{aligned}h_k &:= \operatorname{diam}(\mathrm{T}_k), \quad \mathrm{T}_k \in \mathcal{T}_h,\\ h_\tau &:= \text{length of side } \tau, \quad \tau \in \mathcal{E}_h,\\ h &:= \max_{\mathrm{T}_k \in \mathcal{T}_h} (h_k).\end{aligned}$$

In order to define the weak and finite element solutions to our problem we use some shorthand notation: $\mathcal{V} := \mathrm{H}^1(\Omega)$ and

$$\mathcal{V}_g := \{v \in \mathcal{V} : v = g \text{ on } \partial\Omega_D\} \tag{5.1.4}$$

to denote the natural spaces for the weak solution. The piecewise linear finite element space is given by

$$\mathcal{V}_h := \{v \in \mathcal{V} : v|_{\mathrm{T}_k} \text{ is linear}, \ \forall \mathrm{T}_k \in \mathcal{T}_h\}. \tag{5.1.5}$$

For $d \in \mathrm{H}^{\frac{3}{2}}(\partial\Omega_D)$ we define

$$\mathcal{V}_{h,d} := \{v \in \mathcal{V}_h : \ v(x) = d(x) \text{ for all mesh points } x \in \partial\Omega_D\}. \tag{5.1.6}$$

$\mathcal{V}_{h,g}$ is the natural space to seek the piecewise linear finite element solution to (5.1.1)-(5.1.3).

The weak solution, $u_0 \in \mathcal{V}_g$, of (5.1.1)-(5.1.3) is defined via the functional $\mathrm{F} : \mathcal{V} \to (\mathcal{V}_0)'$, where $(\mathcal{V})'$ is defined to be the dual space of $\mathcal{V}$:

$$(\mathrm{F}(u).v) := (\nabla u, \nabla v) + (f(u), v), \ u \in \mathcal{V}, v \in \mathcal{V}_0. \tag{5.1.7}$$

With this definition the weak solution, $u_0 \in \mathcal{V}_g$, satisfies

$$\mathrm{F}(u_0) = 0 \ \text{ in } (\mathcal{V}_0)'. \tag{5.1.8}$$

We also need the Fréchet Derivative of F. $\mathrm{F}' : \mathcal{V} \to L\left(\mathcal{V}_0, (\mathcal{V}_0)'\right)$, which is defined

by

$$\left(\mathrm{F}'(u)v, w\right) = (\nabla v, \nabla w) + (f'(u)v, w), \ u \in \mathcal{V}, \ v, w \in \mathcal{V}_0. \tag{5.1.9}$$

The finite element approximation, $u_h \in \mathcal{V}_{h,g}$, to u_0 may be defined by the functional $\mathrm{F}_h : \mathcal{V}_h \to (\mathcal{V}_{h,0})'$ given by

$$(\mathrm{F}_h(u), v) := (\nabla u, \nabla v) + (f(u), v), \ u \in \mathcal{V}_h, v \in \mathcal{V}_{h,0}. \tag{5.1.10}$$

Then $u_h \in \mathcal{V}_{h,g}$ satisfies

$$\mathrm{F}_h(u_h) = 0 \ \text{ in } (\mathcal{V}_{h,0})'. \tag{5.1.11}$$

In our *a posteriori* error estimate we need the jump function, $\left[\frac{\partial u_h}{\partial n}\right]_\tau$, of the normal derivative of u_h across an edge τ of a triangle. We define this by

$$\left[\frac{\partial u_h}{\partial n}\right]_\tau := \begin{cases} \nabla u_h|_{\mathrm{T}_-}.n_-(\tau) + \nabla u_h|_{\mathrm{T}_+}.n_+(\tau) & \tau \in \mathcal{E}_\Omega \\ \nabla u_h|_{\mathrm{T}}.n(\tau) & \tau \in \mathcal{E}_N \\ 0 & \tau \in \mathcal{E}_D \end{cases}. \tag{5.1.12}$$

In (5.1.12), if $\tau \in \mathcal{E}_\Omega$: T_+ and T_- are the two triangles common to edge τ with outward normals $n_+(\tau)$ and $n_-(\tau)$ at the edge τ; if $\tau \in \mathcal{E}_N$: T is the triangle on the boundary of Ω to which τ belongs and $n(\tau)$ is its outward normal at the edge. We note that, for piecewise linear u_h, $\left[\frac{\partial u_h}{\partial n}\right]_\tau$ is constant on the edge τ.

We make the following further assumptions on our problem:

(A3) There exists a weak solution, u_0, satisfying (5.1.8). u_0 is a member of the (fractional order) Sobolev space $\mathrm{H}^{1+\alpha} \cap \mathcal{V}_g$, where α is a fixed number, greater than zero, depending only on the geometry of Ω.

(A4) The Fréchet derivative of our weak form, (5.1.9), has the following regularity property at u_0: for all $b \in \mathrm{L}_2$, there exists a unique $w \in \mathrm{H}^{1+\alpha} \cap \mathcal{V}_0$ solving

$$\left(\mathrm{F}'(u_0)w, v\right) = (b, v) \ \ v \in \mathcal{V}_0$$

and furthermore for any constant Λ^0 such that $\|(\mathrm{F}'(u_0))^{-1}\|_{L((\mathcal{V}_0)', \mathcal{V}_0)} \leq \Lambda^0$:

$$\|w\|_{1+\alpha} \leq \Lambda^0 \|b\|_0. \tag{5.1.13}$$

(A5) There exists a finite element solution, $u_h \in \mathcal{V}_{h,g}$, satisfying (5.1.11). We also assume that $\|u_h\|_\infty$ is bounded independently of h, that u_h is locally unique in an H^1-ball centred at u_0 and that

$$\|u_0 - u_h\|_1 \to 0 \text{ as } h \to 0, \tag{5.1.14}$$

$$\|u_0 - u_h\|_\infty \to 0 \text{ as } h \to 0. \tag{5.1.15}$$

Remark 5.1.2 *1. It is shown in Chapter 4 that (A3) follows if we assume there exists a weak solution in L_∞ in addition to (A1) and (A2).*

2. (A5) follows if we assume mesh conditions (M1)-(M5) in Chapter 4.

3. The assumption that $u_0 \in H^{1+\alpha}$ implies that $u_0 \in L_\infty$ by the Sobolev Embedding Theorems.

5.2 The *a posteriori* Error Estimates

In this section we prove the following *a posteriori* error estimates:

$$\|u_0 - u_h\|_1 \leq C\Lambda^0 \left[\left\{ \sum_{\tau \in \mathcal{E}_h} h_\tau^2 \left[\frac{\partial u_h}{\partial n} \right]_\tau^2 \right\}^{\frac{1}{2}} + \left\{ \sum_{T_k \in \mathcal{T}_h} \|h_k f(u_h)\|^2_{L_2(T_k)} \right\}^{\frac{1}{2}} \right] \tag{5.2.16}$$

$$\|u_0 - u_h\|_0 \leq C'\Lambda^0 \left[\left\{ \sum_{T_k \in \mathcal{T}_h} h_k^{2\alpha} \|u_0 - u_h\|^2_{H^1(T_k)} \right\}^{\frac{1}{2}} + \|u_0 - u_h\|_1^2 \right]. \tag{5.2.17}$$

In the above C and C' are constants independent of f, the mesh parameters and the finite element solution and Λ^0 is the bound on the inverse of $F'(u_0)$ appearing in (5.1.13).

These *a posteriori* error estimates are analogous to the estimates in [70] and [71]. However, in [71], the loss of H^2 regularity due to reentrant corners and/or mixed boundary conditions is handled by use of a scale of W_p^1 spaces with variable p. In this work we instead use the scale $H^{1+\alpha} = W_2^{1+\alpha}$. Similar L_2 estimates, but assuming full H^2 regularity, are found in [31].

As we are mainly interested in these error estimates for mesh refinement, we do not calculate the value of the constants C, C' and Λ^0 here, but instead estimate their values

numerically as part of our mesh refinement strategy.

At the end of this section we also give the *a posteriori* error estimates for a one dimension semilinear problem. The results are somewhat surprising as the estimate does not contain a term derived from the Laplacian.

5.2.1 The H^1 Estimate

Here we prove the following theorem which gives the *a posteriori* error estimate in the H^1-norm:

Theorem 5.2.1 *Let $u_0 \in \mathcal{V}_g$ be the solution to problem (5.1.8) and let $u_h \in \mathcal{V}_{h,g}$ be the solution of (5.1.11). Then if h is sufficiently small:*

$$\|u_0 - u_h\|_1 \leq C\Lambda^0 \left[\left\{ \sum_{\tau \in \mathcal{E}_h} h_\tau^2 \left[\frac{\partial u_h}{\partial n} \right]_\tau^2 \right\}^{\frac{1}{2}} + \left\{ \sum_{\mathrm{T}_k \in \mathcal{T}_h} \|h_k f(u_h)\|^2_{\mathrm{L}_2(\mathrm{T}_k)} \right\}^{\frac{1}{2}} \right]. \quad (5.2.18)$$

The proof of this theorem is obtained with the help of Lemma C.1.2 in Appendix C, which shows that providing u_h is sufficiently close to u_0 in the H^1-norm (which we can guarantee, by (A5), if h is small enough):

$$\|u_0 - u_h\|_1 \leq 2\Lambda^0 \|\mathrm{F}(u_h)\|_{(\mathcal{V}_0)'}.$$

We also use the "quasi-interpolant" Verfürth introduces in his paper [70]. This tells us that for each $v \in H^1$, there exists a piecewise linear "quasi-interpolant" $\tilde{v}$ such that

(a) For a triangle T_k, with diameter h_k

$$\|v - \tilde{v}\|_{\mathrm{L}_2(\mathrm{T}_k)} \leq C h_k \|v\|_{\mathrm{H}^1(\tilde{\mathrm{T}}_k)} \quad (5.2.19)$$

where $\tilde{\mathrm{T}}_k$ is the union of all triangles in $\mathcal{T}_h$ having a non-empty intersection with T_k.

(b) For an edge τ triangle in the triangulation, with length h_τ.

$$\|v - \tilde{v}\|_{\mathrm{L}_2(\tau)} \leq C(h_\tau)^{\frac{1}{2}} \|v\|_{\mathrm{H}^1(\tilde{\tau})} \quad (5.2.20)$$

where $\tilde{\tau}$ is the union of all triangles in $\mathcal{T}_h$ having a non-empty intersection with τ.

To prove the estimates (5.2.19) and (5.2.20) Verfürth uses results by Clément [21] together with a scaling argument. Similar estimates are also proved in [13].

We are now in a position to prove Theorem 5.2.1:

Proof From assumptions (A3) and (A5) we have that u_0 and u_h are bounded (independently of h) and that $\|u_0 - u_h\|_1 \to 0$ as $h \to 0$.

Thus, for small enough h, we may apply Lemma C.1.2 of Appendix C to prove that for the functional F given by (5.1.7),

$$\|u_0 - u_h\|_1 \leq 2\Lambda^0 \|\mathrm{F}(u_h)\|_{(\mathcal{V}_0)'}. \tag{5.2.21}$$

In (5.2.21) Λ^0 is the bound on the inverse of $\mathrm{F}'(u_0)$ introduced in assumption (A4).

Thus it remains to bound $\|\mathrm{F}(u_h)\|_{(\mathcal{V}_0)'}$. By definition

$$\|\mathrm{F}(u_h)\|_{(\mathcal{V}_0)'} = \sup_{v \in \mathcal{V}_0,\ \|v\|_1 = 1} |(\mathrm{F}(u_h), v)|. \tag{5.2.22}$$

Thus we aim to prove the theorem by bounding $|(\mathrm{F}(u_h), v)|$ above in terms of u_h and v.

Take any $v \in \mathcal{V}_0$ and let $\tilde{v} \in \mathcal{V}_{h,0}$ be its "quasi-interpolant" satisfying (5.2.19) and (5.2.20). Then since u_h satisfies (5.1.11):

$$\begin{aligned}
(\mathrm{F}(u_h), v) &= (\nabla u_h, \nabla v) + (f(u_h), v) \\
&= (\nabla u_h, \nabla[v - \tilde{v}]) + (f(u_h), [v - \tilde{v}]) \\
&=: S_1 + S_2.
\end{aligned} \tag{5.2.23}$$

We bound S_1 and S_2 separately.

First we consider S_1 and use Green's Theorem in each triangle, T_k:

$$\begin{aligned}
S_1 &= (\nabla u_h, \nabla[v - \tilde{v}]) \\
&= \sum_{\mathrm{T}_k \in \mathcal{T}_h} (\nabla u_h, \nabla[v - \tilde{v}])_{\mathrm{T}_k} \\
&= \sum_{\mathrm{T}_k \in \mathcal{T}_h} \left[\sum_{\tau \in \mathcal{E}(\mathrm{T}_k)} \int_\tau \frac{\partial u_h}{\partial n_\tau}(v - \tilde{v}) - \int_{\mathrm{T}_k} (\Delta u_h)(v - \tilde{v}) \right] \\
&= \sum_{\mathrm{T}_k \in \mathcal{T}_h} \sum_{\tau \in \mathcal{E}(\mathrm{T}_k)} \int_\tau \frac{\partial u_h}{\partial n_\tau}(v - \tilde{v}).
\end{aligned}$$

where the last step follows since $\Delta u_h|_{\mathrm{T}_k} \equiv 0$ for all $\mathrm{T}_k \in \mathcal{T}_h$.

Now, taking the sum over all the edges of the triangles and remembering that $v - \tilde{v} = 0$ on $\partial\Omega_D$ ($v \in \mathcal{V}_0$, $\tilde{v}$ is piecewise linear and zero at all the mesh points on $\partial\Omega_D$) we have that:

$$\begin{aligned} |S_1| &= |\sum_{\mathrm{T}_k \in \mathcal{T}_h} \sum_{\tau \in \mathcal{E}(\mathrm{T}_k)} \int_\tau \frac{\partial u_h}{\partial n_\tau}(v - \tilde{v})| \\ &\leq \sum_{\tau \in \mathcal{E}_h} |\int_\tau \left[\frac{\partial u_h}{\partial n}\right]_\tau (v - \tilde{v})| \\ &\leq \sum_{\tau \in \mathcal{E}_h} |\int_\tau (h_\tau)^{\frac{1}{2}} \left[\frac{\partial u_h}{\partial n}\right]_\tau (h_\tau)^{-\frac{1}{2}}(v - \tilde{v})|. \end{aligned}$$

Now using the Cauchy-Schwartz inequality twice, the fact that $\left[\frac{\partial u_h}{\partial n}\right]_\tau$ is constant on each edge $\tau \in \mathcal{E}_h$ and (5.2.20), we have the chain of inequalities

$$\begin{aligned} |S_1| &\leq \sum_{\tau \in \mathcal{E}_h} \left\{ \int_\tau \left((h_\tau)^{\frac{1}{2}} \left[\frac{\partial u_h}{\partial n}\right]_\tau \right)^2 \right\}^{\frac{1}{2}} \left\{ \int_\tau \left((h_\tau)^{-\frac{1}{2}}(v - \tilde{v}) \right)^2 \right\}^{\frac{1}{2}} \\ &= \sum_{\tau \in \mathcal{E}_h} \|(h_\tau)^{\frac{1}{2}} \left[\frac{\partial u_h}{\partial n}\right]_\tau \|_{\mathrm{L}_2(\tau)} \|(h_\tau)^{-\frac{1}{2}}(v - \tilde{v})\|_{\mathrm{L}_2(\tau)} \\ &\leq \left\{ \sum_{\tau \in \mathcal{E}_h} \|(h_\tau)^{\frac{1}{2}} \left[\frac{\partial u_h}{\partial n}\right]_\tau \|^2_{\mathrm{L}_2(\tau)} \right\}^{\frac{1}{2}} \left\{ \sum_{\tau \in \mathcal{E}_h} \|(h_\tau)^{-\frac{1}{2}}(v - \tilde{v})\|^2_{\mathrm{L}_2(\tau)} \right\}^{\frac{1}{2}} \\ &\leq C \left\{ \sum_{\tau \in \mathcal{E}_h} (h_\tau)^2 \left[\frac{\partial u_h}{\partial n}\right]^2_\tau \right\}^{\frac{1}{2}} \left\{ \sum_{\tau \in \mathcal{E}_h} \|v\|^2_{\mathrm{H}^1(\tilde{\tau})} \right\}^{\frac{1}{2}} \\ &\leq C \left\{ \sum_{\tau \in \mathcal{E}_h} (h_\tau)^2 \left[\frac{\partial u_h}{\partial n}\right]^2_\tau \right\}^{\frac{1}{2}} \|v\|_1. \end{aligned} \tag{5.2.24}$$

The constant in the last line depends on the maximum number of triangles which touch any given triangle. This is assumed bounded as $h \to 0$ in assumption (M4).

Now we go on to consider bounding S_2 above.

$$\begin{aligned} |S_2| &= |(f(u_h), [v - \tilde{v}])| \\ &= |\sum_{\mathrm{T}_k \in \mathcal{T}_h} (f(u_h), [v - \tilde{v}])_{\mathrm{T}_k}| \\ &\leq \sum_{\mathrm{T}_k \in \mathcal{T}_h} |(h_k\, f(u_h), (h_k)^{-1}[v - \tilde{v}])_{\mathrm{T}_k}|. \end{aligned}$$

Then, using the Cauchy-Schwartz inequality twice and (5.2.19), we obtain

$$\begin{aligned}
|S_2| &\leq \sum_{\mathrm{T}_k \in \mathcal{T}_h} \left\{ \int_{\mathrm{T}_k} (h_k \, f(u_h))^2 \right\}^{\frac{1}{2}} \left\{ \int_{\mathrm{T}_k} \left((h_k)^{-1} [v - \tilde{v}] \right)^2 \right\}^{\frac{1}{2}} \\
&= \left\{ \sum_{\mathrm{T}_k \in \mathcal{T}_h} \| h_k \, f(u_h) \|^2_{\mathrm{L}_2(\mathrm{T}_k)} \right\}^{\frac{1}{2}} \left\{ \sum_{\mathrm{T}_k} \| (h_k)^{-1} [v - \tilde{v}] \|^2_{\mathrm{L}_2(\mathrm{T}_k)} \right\}^{\frac{1}{2}} \\
&\leq C \left\{ \sum_{\mathrm{T}_k \in \mathcal{T}_h} \| h_k \, f(u_h) \|^2_{\mathrm{L}_2(\mathrm{T}_k)} \right\}^{\frac{1}{2}} \left\{ \sum_{\mathrm{T}_k} \| v \|^2_{\mathrm{H}^1(\tilde{\mathrm{T}}_k)} \right\}^{\frac{1}{2}} \\
&\leq C \left\{ \sum_{\mathrm{T}_k \in \mathcal{T}_h} \| h_k \, f(u_h) \|^2_{\mathrm{L}_2(\mathrm{T}_k)} \right\}^{\frac{1}{2}} \| v \|_1, \qquad (5.2.25)
\end{aligned}$$

with the constant C depending on the connectivity of the mesh as in (5.2.24).

Thus combining (5.2.24) and (5.2.25) on S_1 with (5.2.23) and (5.2.22) we obtain the result

$$\| \mathrm{F}(u_h) \|_{(\mathcal{V}_0)'} \leq C \left[\left\{ \sum_{\tau \in \mathcal{E}_h} (h_\tau)^2 \left[\frac{\partial u_h}{\partial n} \right]^2_\tau \right\}^{\frac{1}{2}} + \left\{ \sum_{\mathrm{T}_k \in \mathcal{T}_h} \| h_k \, f(u_h) \|^2_{\mathrm{L}_2(\mathrm{T}_k)} \right\}^{\frac{1}{2}} \right].$$

Combining this bound with (5.2.21) proves the lemma. □

5.2.2 The L_2 Estimate

In this section we prove an *a posteriori* error estimate in the L_2-norm. This is proved using a duality argument which takes us from the H^1 estimate to the L_2 estimate. For technical reasons we assume here that g in (5.1.1)-(5.1.3) is replaced by its piecewise linear interpolant on $\partial\Omega_D$. This is what is usually done in practice and allows us to write a simpler proof of the theorem. This result could be extended to include the case of general g by the techniques used in Lemma 4.2.9. The result is contained in the following theorem:

Theorem 5.2.2 *Let $u_0 \in \mathcal{V}_g$ be the solution to problem (5.1.8) and let $u_h \in \mathcal{V}_{h,g}$ be the solution of (5.1.11). Then if h is sufficiently small:*

$$\| u_0 - u_h \|_0 \leq C \Lambda^0 \left[\left\{ \sum_{\mathrm{T}_k \in \mathcal{T}_h} (h_k)^{2\alpha} \| u_0 - u_h \|^2_{\mathrm{H}^1(\mathrm{T}_k)} \right\}^{\frac{1}{2}} + \| u_0 - u_h \|^2_1 \right]. \qquad (5.2.26)$$

To prove this theorem we need an interpolation operator which is valid for Sobolev spaces of fractional order. We use the interpolation operator and estimates proved in Scott and Zhang [61]. This operator is different to the interpolant used in Section 5.2.1. The paper states that for all $v \in \mathcal{V}_0$, there exists a $\hat{v} \in \mathcal{V}_{h,0}$ such that for $0 \leq m \leq l \leq 2$, $l > \frac{1}{2}$:

$$\left\{ \sum_{\mathrm{T}_k \in \mathcal{T}_h} (h_k)^{2(m-l)} \|v - \hat{v}\|^2_{\mathrm{H}^m(\mathrm{T}_k)} \right\}^{\frac{1}{2}} \leq C\|v\|_l, \tag{5.2.27}$$

which implies

$$\|v - \hat{v}\|_m \leq Ch^{l-m}\|v\|_l. \tag{5.2.28}$$

We can now give the proof of Theorem 5.2.2:

Proof Define $e_h = u_0 - u_h$, $e_h \in \mathcal{V}_0$ by assumption. We consider the auxiliary problem:

Seek $\chi \in \mathcal{V}_0$ such that

$$\begin{aligned} -\Delta\chi + f'(u_0)\chi &= e_h, \quad \text{in } \Omega, && (5.2.29)\\ \chi &= 0, \quad \text{on } \partial\Omega_D, && (5.2.30)\\ \frac{\partial\chi}{\partial n} &= 0, \quad \text{on } \partial\Omega_N. && (5.2.31) \end{aligned}$$

The weak solution, χ_0, to (5.2.29)-(5.2.31), solves the following problem:

Seek $\chi_0 \in \mathcal{V}_0$ such that

$$(\nabla\chi_0, \nabla v) + (f'(u_0)\chi_0, v) = (e_h, v), \quad v \in \mathcal{V}_0 \tag{5.2.32}$$

or equivalently seek $\chi_0 \in \mathcal{V}_0$ such that:

$$\mathrm{F}'(u_0)\chi_0 = e_h \ \text{ in } (\mathcal{V}_0)'.$$

By the assumption (A4), χ_0 is in the space $\mathrm{H}^{1+\alpha} \cap \mathcal{V}_0$ and

$$\|\chi_0\|_1 \leq \|\chi_0\|_{1+\alpha} \leq \Lambda^0\|e_h\|_0. \tag{5.2.33}$$

Using (5.2.32) we have that

$$\|e_h\|_0^2 = (e_h, e_h) = (\nabla\chi_0, \nabla e_h) + (f'(u_0)\chi_0, e_h). \tag{5.2.34}$$

Taking $\hat{\chi}_0 \in \mathcal{V}_{h,0}$ to be the interpolant defined in [61] and satisfying (5.2.27) and (5.2.28) we note that

$$(\nabla u_0, \nabla\hat{\chi}_0) + (f(u_0), \hat{\chi}_0) = 0$$

and

$$(\nabla u_h, \nabla\hat{\chi}_0) + (f(u_h), \hat{\chi}_0) = 0.$$

Thus

$$(\nabla e_h, \nabla\hat{\chi}_0) + (f(u_0) - f(u_h), \hat{\chi}_0) = 0. \tag{5.2.35}$$

So taking (5.2.35) from (5.2.34) we obtain

$$\begin{aligned}
\|e_h\|_0^2 &= (\nabla e_h, \nabla(\chi_0 - \hat{\chi}_0)) - (f(u_0) - f(u_h), \hat{\chi}_0) + (f'(u_0)\chi_0, e_h) \\
&= (\nabla e_h, \nabla(\chi_0 - \hat{\chi}_0)) - (f(u_0) - f(u_h) - f'(u_0)e_h, \hat{\chi}_0) + (f'(u_0)e_h, \chi_0 - \hat{\chi}_0) \\
&=: S_1 + S_2 + S_3. \qquad (5.2.36)
\end{aligned}$$

We bound S_1, S_2 and S_3 separately.

First we consider S_1. Using the Cauchy-Schwartz inequality twice we have

$$\begin{aligned}
|S_1| &= |(\nabla e_h, \nabla(\chi_0 - \hat{\chi}_0))| \\
&= |\sum_{T_k\in\mathcal{T}_h} (\nabla e_h, \nabla(\chi_0 - \hat{\chi}_0))_{T_k}| \\
&= |\sum_{T_k\in\mathcal{T}_h} ((h_k)^\alpha \nabla e_h, (h_k)^{-\alpha}\nabla(\chi_0 - \hat{\chi}_0))_{T_k}| \\
&\leq \left\{\sum_{T_k\in\mathcal{T}_h} \|(h_k)^\alpha \nabla e_h\|^2_{L_2(T_k)}\right\}^{\frac{1}{2}} \left\{\sum_{T_k\in\mathcal{T}_h} \|(h_k)^{-\alpha}\nabla(\chi_0 - \hat{\chi}_0)\|^2_{L_2(T_k)}\right\}^{\frac{1}{2}}.
\end{aligned}$$

Now, using (5.2.27) with $m = 1$ and $l = 1 + \alpha$, we have the result that

$$|S_1| \leq C\left\{\sum_{T_k\in\mathcal{T}_h} \|(h_k)^\alpha \nabla e_h\|^2_{L_2(T_k)}\right\}^{\frac{1}{2}} \|\chi_0\|_{1+\alpha}$$

and using (5.2.33) we deduce that

$$\begin{aligned} |S_1| &\leq C\Lambda^0 \left\{ \sum_{T_k \in \mathcal{T}_h} \|(h_k)^{\alpha} \nabla e_h\|^2_{L_2(T_k)} \right\}^{\frac{1}{2}} \|e_h\|_0 \\ &\leq C\Lambda^0 \left\{ \sum_{T_k \in \mathcal{T}_h} \|(h_k)^{\alpha} e_h\|^2_{H^1(T_k)} \right\}^{\frac{1}{2}} \|e_h\|_0. \end{aligned} \tag{5.2.37}$$

Now we consider bounding S_3:

$$\begin{aligned} |S_3| &= |(f'(u_0)e_h, \chi_0 - \hat{\chi}_0)| \\ &= |\sum_{T_k \in \mathcal{T}_h} (f'(u_0)e_h, \chi_0 - \hat{\chi}_0)_{T_k}| \\ &\leq C \sum_{T_k \in \mathcal{T}_h} ((h_k)^{(1+\alpha)}|e_h|, (h_k)^{-(1+\alpha)}|\chi_0 - \hat{\chi}_0|)_{T_k} \\ &\leq C \left\{ \sum_{T_k \in \mathcal{T}_h} \|(h_k)^{(1+\alpha)} e_h\|^2_{L_2(T_k)} \right\}^{\frac{1}{2}} \left\{ \sum_{T_k \in \mathcal{T}_h} \|(h_k)^{-(1+\alpha)}(\chi_0 - \hat{\chi}_0)\|^2_{L_2(T_k)} \right\}^{\frac{1}{2}} \end{aligned}$$

here $C = \|f'(u_0)\|_\infty$.

Again we return to (5.2.27), with $m = 0$ and $l = 1 + \alpha$, and (5.2.33) to obtain the inequality:

$$\begin{aligned} |S_3| &\leq C \left\{ \sum_{T_k \in \mathcal{T}_h} \|(h_k)^{(1+\alpha)} e_h\|^2_{L_2(T_k)} \right\}^{\frac{1}{2}} \|\chi_0\|_{1+\alpha} \\ &\leq C\Lambda^0 \left\{ \sum_{T_k \in \mathcal{T}_h} \|(h_k)^{(1+\alpha)} e_h\|^2_{L_2(T_k)} \right\}^{\frac{1}{2}} \|e_h\|_0 \\ &\leq C\Lambda^0 \left\{ \sum_{T_k \in \mathcal{T}_h} \|(h_k)^{1+\alpha} e_h\|^2_{H^1(T_k)} \right\}^{\frac{1}{2}} \|e_h\|_0 \\ &\leq C\Lambda^0 \left\{ \sum_{T_k \in \mathcal{T}_h} \|(h_k)^{\alpha} e_h\|^2_{H^1(T_k)} \right\}^{\frac{1}{2}} \|e_h\|_0. \end{aligned} \tag{5.2.38}$$

which is the same as the bound on $|S_1|$.

Before bounding S_2, we note that by Taylor's theorem, for each $\boldsymbol{x} \in \Omega$

$$f(u_h(\boldsymbol{x})) = f(u_0(\boldsymbol{x})) + f'(u_0(\boldsymbol{x}))(u_h(\boldsymbol{x}) - u_0(\boldsymbol{x}))$$

$$+ f''(\sigma(\boldsymbol{x}))(u_h(\boldsymbol{x}) - u_0(\boldsymbol{x}))^2$$

where $\sigma(\boldsymbol{x})$ lies between $u_0(\boldsymbol{x})$ and $u_h(\boldsymbol{x})$. Thus

$$\begin{aligned} |\sigma(\boldsymbol{x})| &\leq \max\{|u_0(\boldsymbol{x})|, |u_h(\boldsymbol{x})|\} \\ &\leq \gamma \text{ independent of } h \text{ by assumption (A5).} \end{aligned}$$

Hence

$$|f(u_0) - f(u_h) - f'(u_0)e_h| \leq \|f''\|_{\mathrm{L}_\infty[-\gamma,\gamma]} |e_h|^2. \tag{5.2.39}$$

Thus, we can conclude that

$$\begin{aligned} |S_2| &= |(f(u_0) - f(u_h) - f'(u_0)e_h, \hat{\chi}_0)| \\ &\leq C(|e_h|^2, |\hat{\chi}_0|). \end{aligned}$$

Now by the Generalised Hölder's inequality and (5.2.28) with $m = 1$ and $l = 1 + \alpha$ we have that

$$\begin{aligned} |S_2| &\leq C(|e_h|^2, |\hat{\chi}_0|) \\ &\leq C\|e_h\|_{\mathrm{L}_4}^2 \, \|\hat{\chi}_0\|_0 \\ &\leq C\|e_h\|_1^2 \, \|\chi_0\|_1 \\ &\leq C\|e_h\|_1^2 \, [\|\chi_0 - \hat{\chi}_0\|_1 + \|\chi_0\|_1] \\ &\leq C\|e_h\|_1^2 \, [1 + h^\alpha] \, \|\chi_0\|_{1+\alpha}. \end{aligned}$$

Now using (5.2.33), we obtain the result

$$\begin{aligned} |S_2| &\leq C\|e_h\|_1^2 \, [1 + h^\alpha] \, \Lambda^0 \|e_h\|_0 \\ &\leq C\Lambda^0 \|e_h\|_1^2 \, \|e_h\|_0. \end{aligned} \tag{5.2.40}$$

Thus, returning to (5.2.36) and using (5.2.37), (5.2.40) and (5.2.38), we obtain the required result. □

In order to implement the L_2 estimate in our adaptive procedure we need to know the H^1 *a posteriori* error estimate on a triangle. We estimate this quantity by the contribution a triangle, T_k, makes to the overall H^1 *a posteriori* error estimate, i.e. we

shall assume that we have the following estimate:

$$\|u_0 - u_h\|_{\mathrm{H}^1(\mathrm{T}_k)} \leq C\Lambda^0 \left[\left\{ \sum_{\tau \in \mathcal{E}(\mathrm{T}_k)} (h_\tau)^2 \left[\frac{\partial u_h}{\partial n} \right]_\tau^2 \right\}^{\frac{1}{2}} + \|h_k f(u_h)\|_{\mathrm{L}_2(\mathrm{T}_k)} \right]. \quad (5.2.41)$$

5.2.3 The *a posteriori* Error Estimate in One Dimension

We consider finding the H^1 and L_2 *a posteriori* error estimates for the following one dimensional semilinear problem:

Find u such that:

$$-u''(x) + f(u(x), x) = 0, \qquad \text{on } \Omega = [0, 1], \quad (5.2.42)$$

$$u(0) = u^0, \qquad u(1) = u^1. \quad (5.2.43)$$

where u^0 and u^1 are given and $f : \mathbb{R} \times \Omega \to \mathbb{R}$ has the property that for all $x \in \Omega$, $f(\cdot, x) \in \mathcal{C}^2(\mathbb{R})$ and if $u \in \mathcal{C}(\Omega)$ then the function $x \to f(u(x), x)$ is in $\mathrm{L}_\infty(\Omega)$.

We define $u_0 \in X := \left\{ v \in \mathrm{H}^1(\Omega) : v(0) = u^0, v(1) = u^1 \right\}$ to be the weak solution of (5.2.42), (5.2.43) which we assume to exist. Then u_0 solves the problem

$$\mathrm{F}(u) = 0 \ \text{ in } \left(\mathrm{H}_0^1\right)', \quad (5.2.44)$$

where $\mathrm{F} : X \to \left(\mathrm{H}_0^1\right)'$ is defined by

$$(\mathrm{F}(u), v) := (u', v') + (f(u), v), \quad u \in X, v \in \mathrm{H}_0^1. \quad (5.2.45)$$

We also define the Fréchet Derivative of (5.2.45) by $\mathrm{F}' : X \to L(\mathrm{H}_0^1, \left(\mathrm{H}_0^1\right)')$

$$(\mathrm{F}'(u)v, w) := (v', w') + (f'(u)v, w), \quad u \in X, \ v, w \in \mathrm{H}_0^1. \quad (5.2.46)$$

We seek a finite element approximation to (5.2.42), (5.2.43). To do this we define a finite element mesh with $n + 2$ mesh points, $\{x_k\}_0^{n+1}$, such that

$$0 = x_0 < x_1 < \ldots < x_n < x_{n+1} = 1.$$

Define the mesh parameters:

$$\begin{aligned} h_k &:= x_k - x_{k-1}, \quad k = 1, \ldots, n+1, \\ h &:= \max_{k=1,\ldots,n+1} \{h_k\}, \\ I_k &:= [x_{k-1}, x_k], \quad k = 1, \ldots, n+1. \end{aligned}$$

Define $X_h := \left\{v \in X : v(0) = u^0, v(1) = u^1, v|_{I_k} \text{ is linear, } k = 1, \ldots, n+1\right\}$. We consider finding the piecewise linear finite element approximation, $u_h \in X_h$, of the weak solution of (5.2.42), (5.2.43). Introducing the functional $\mathrm{F}_h : X_h \to (X_{h,0})'$:

$$(\mathrm{F}_h(u), v) := (u', v') + (f(u), v), \quad u \in X_h, v \in X_{h,0}, \tag{5.2.47}$$

where $X_{h,0} := \{v \in X_h : v(0) = v(1) = 0\}$, we require u_h to solve:

$$\mathrm{F}_h(u_h) = 0 \text{ in } (X_{h,0})'. \tag{5.2.48}$$

We make the following assumptions on our one dimensional semilinear problem and its finite element approximation:

(A6) There exist a weak solution, $u_0 \in X \cap \mathrm{H}^2(\Omega)$, satisfying (5.2.44).

(A7) There exists a finite element solution, $u_h \in X_h$, satisfying (5.2.48) which is locally unique in an H^1-ball centred at u_0 and

$$\begin{aligned} \|u_0 - u_h\|_1 &\to 0 \quad \text{as} \quad h \to 0, \\ \|u_0 - u_h\|_\infty &\to 0 \quad \text{as} \quad h \to 0. \end{aligned}$$

(A8) $\mathrm{F}'(u_0) : \mathrm{H}_0^1 \to (\mathrm{H}_0^1)'$ is a bounded invertible functional and for all $b \in \mathrm{L}_2$, there exists a unique $w \in \mathrm{H}^2 \cap \mathrm{H}_0^1$ solving

$$\mathrm{F}'(u_0)w = b \text{ in } (\mathrm{H}_0^1)'$$

and

$$\|w\|_2 \leq \Lambda_0 \|b\|_0$$

where $\| \left(\mathrm{F}'(u_0)\right)^{-1} \|_{L((\mathrm{H}^1)', \mathrm{H}^1)} \leq \Lambda^0$.

With the assumptions (A6)-(A8) we outline the proof of the following one dimensional *a posteriori* error estimates:

$$\begin{aligned}\|u_0 - u_h\|_1 &\leq C\Lambda_0 \left\{ \sum_{k=1}^{n+1} \|h_k f(u_h)\|^2_{\mathrm{L}_2(I_k)} \right\}^{\frac{1}{2}}, \\ \|u_0 - u_h\|_0 &\leq C\Lambda_0 \left\{ \left\{ \sum_{k=1}^{n+1} h_k^2 \|u_0 - u_h\|^2_{\mathrm{H}^1(I_k)} \right\}^{\frac{1}{2}} + \|u_0 - u_h\|_1^2 \right\}.\end{aligned}$$

To prove these error estimates we need the standard finite element interpolant:

Let v be a member of H_0^1, then v is continuous and there exists a $\tilde{v} \in X_{h,0}$ such that for each mesh point x_k:

$$v(x_k) = \tilde{v}(x_k), \quad k = 0, \ldots, n+1.$$

It is well known, see for example [42], that

$$\|v - \tilde{v}\|_{\mathrm{L}_2(I_k)} \leq Ch_k \|v'\|_{\mathrm{L}_2(I_k)}, \tag{5.2.49}$$

$$\|v - \tilde{v}\|_{\mathrm{L}_2(I_k)} \leq Ch_k^2 \|v''\|_{\mathrm{L}_2(I_k)}. \tag{5.2.50}$$

We also need the following result, which is proved in an identical way to the corresponding result (5.2.21) in the two dimension case (by using assumptions (A6)-(A8) and a one dimensional version of Lemmas C.1.1 and C.1.2 in Appendix C):

For h sufficiently small:

$$\|u_0 - u_h\|_1 \leq 2\Lambda_0 \|\mathrm{F}(u_h)\|_{(X)'}. \tag{5.2.51}$$

We are now able to we outline the proof of the H^1 *a posteriori* error estimates in one dimension:

Theorem 5.2.3 *Let $u_0 \in X$ be the solution to problem (5.2.44) and $u_h \in X_h$ solve the problem (5.2.48), then for h sufficiently small:*

$$\|u_0 - u_h\|_1 \leq C\Lambda_0 \left\{ \sum_{k=1}^{n+1} \|h_k f(u_h)\|^2_{\mathrm{L}_2(I_k)} \right\}^{\frac{1}{2}}. \tag{5.2.52}$$

Proof We only outline the proof as it is very similar to the proof of Theorem 5.2.1.

For arbitrary $v \in X_0$ define $\tilde{v} \in X_{h,0}$ to be the interpolant satisfying (5.2.49) and (5.2.50). Then:

$$\begin{aligned}(\mathrm{F}(u_h), v) &= (u_h', v') + (f(u_h), v) \\ &= (u_h', (v - \tilde{v})') + (f(u_h), (v - \tilde{v})).\end{aligned}$$

Then using integration by parts over each interval, I_k, and remembering that for a mesh point, x_k,

$$(v - \tilde{v})(x_k) = 0$$

and since u_h is a piecewise linear function $u_h''|_{I_k} \equiv 0$, we obtain

$$(\mathrm{F}(u_h), v) = \sum_{k=1}^{n+1} (f(u_h), (v - \tilde{v}))_{I_k}.$$

Thus, using the Cauchy-Schwartz inequality twice:

$$\begin{aligned}|(\mathrm{F}(u_h), v)| &\leq \sum_{k=1}^{n+1} |(h_k f(u_h), (h_k)^{-1}(v - \tilde{v}))_{I_k}| \\ &\leq \left\{ \sum_{k=1}^{n+1} \|h_k f(u_h)\|^2_{\mathrm{L}_2(I_k)} \right\}^{\frac{1}{2}} \left\{ \sum_{k=1}^{n+1} \|(h_k)^{-1}(v - \tilde{v})\|^2_{\mathrm{L}_2(I_k)} \right\}^{\frac{1}{2}}.\end{aligned}$$

Using the interpolation error estimate (5.2.49):

$$\begin{aligned}|(\mathrm{F}(u_h), v)| &\leq C \left\{ \sum_{k=1}^{n+1} \|h_k f(u_h)\|^2_{\mathrm{L}_2(I_k)} \right\}^{\frac{1}{2}} \left\{ \sum_{k=1}^{n+1} \|v'\|^2_{\mathrm{L}_2(I_k)} \right\}^{\frac{1}{2}} \\ &\leq C \left\{ \sum_{k=1}^{n+1} \|h_k f(u_h)\|^2_{\mathrm{L}_2(I_k)} \right\}^{\frac{1}{2}} \|v\|_1.\end{aligned}$$

Therefore

$$\sup_{v \in X_0, \|v\|_1 = 1} |(\mathrm{F}(u_h), v)| \leq C \left\{ \sum_{k=1}^{n+1} \|h_k f(u_h)\|^2_{\mathrm{L}_2(I_k)} \right\}^{\frac{1}{2}}$$

which, when combined with the estimate (5.2.51), implies the result. □

Theorem 5.2.4 *Let $u_0 \in X$ be the solution to problem (5.2.44) and $u_h \in X_h$ solve the*

problem (5.2.48), then for h sufficiently small:

$$\|u_0 - u_h\|_0 \leq C\Lambda_0 \left\{ \left\{ \sum_{k=1}^{n+1} h_k^2 \|u_0 - u_h\|^2_{\mathrm{H}^1(I_k)} \right\}^{\frac{1}{2}} + \|u_0 - u_h\|_1^2 \right\}. \tag{5.2.53}$$

Proof Again we only outline the proof, as it is very similar to the proof of Theorem 5.2.2

Here we define $e_h := u_0 - u_h$ and consider the auxiliary problem:

Seek $z \in \mathrm{H}_0^1$ such that:

$$-z'' + f'(u_0)z = e_h \quad \text{on } \Omega \tag{5.2.54}$$

where u_0 solves (5.2.44).

Let z_0 be the weak solution of (5.2.54), then z_0 solves

$$\mathrm{F}'(u_0)z = e_h \quad \text{in } \left(\mathrm{H}_0^1\right)'$$

and by assumption (A8):

$$\|z_0\|_0 \leq \|z_0\|_1 \leq \|z_0\|_2 \leq \Lambda_0 \|e_h\|_0. \tag{5.2.55}$$

As before define $\tilde{z}_0 \in X_{h,0}$ to be the interpolant of z_0 at the mesh points, satisfying (5.2.49) and (5.2.50). Then as in the two dimension case:

$$\begin{aligned}
\|e_h\|_0^2 &= (e_h, e_h) \\
&= (z_0', e_h') + (f'(u_0)z_0, e_h) \\
&= (e_h', (z_0 - \tilde{z}_0)') + ((f(u_0) - f(u_h)), \tilde{z}_0) + (f'(u_0)e_h, z_0) \\
&= (e_h', (z_0 - \tilde{z}_0)') + ((f(u_0) - f(u_h) + f'(u_0)e_h), \tilde{z}_0) + (f'(u_0)e_h, (z_0 - \tilde{z}_0)) \\
&=: S_1 + S_2 + S_3.
\end{aligned} \tag{5.2.56}$$

Bounding S_1, S_2 and S_3 using (5.2.49), (5.2.50) and (5.2.55), we obtain:

$$\begin{aligned}
|S_1| &\leq C\Lambda_0 \left\{ \sum_{k=1}^{n+1} h_k^2 \|e_h\|^2_{\mathrm{H}^1(I_k)} \right\}^{\frac{1}{2}} \|e_h\|_0, \\
|S_2| &\leq C\Lambda_0 \|e_h\|_1^2 \|e_h\|_0,
\end{aligned}$$

$$|S_3| \leq C\Lambda_0 \left\{ \sum_{k=1}^{n+1} h_k^2 \|e_h\|^2_{\mathrm{H}^1(I_k)} \right\}^{\frac{1}{2}} \|e_h\|_0.$$

Using these bounds, in conjunction with (5.2.56), gives the result. □

Remark 5.2.5 *Although the H^1 a posteriori error estimate (5.2.52) is slightly different to the corresponding two dimensional estimate numerical experiments confirm that error estimates of this form are accurate in the one dimensional case.*

5.3 Adaptive Techniques

There are many different ways of refining a given triangulation (for a nice summary see [53]). We focus on a method that tries to equidistribute the error in the finite element solution over the triangles and produces a conforming triangulation without zero limiting angles.

Our approach in this thesis is as follows:

Assuming the overall *a posteriori* error estimate is larger than a given tolerance, we shall calculate the contribution to the *a posteriori* error estimate (on the whole domain) made by each triangle (see (5.2.41) for the H^1 *a posteriori* error estimate on a triangle). We then need to identify a list of triangles whose contribution to the error is large compared to the average error on a triangle (these triangles are to be refined). We also form a list of triangles where the error is very small (these triangles are to be derefined, if possible). In fact we identify a triangle for refinement if

$$\text{contribution to the error} > 2 \times \text{average contribution to the error}$$

Furthermore we require that we refine a maximum of 300 triangles at each iteration of the algorithm. If we have identified more than 300 triangles we increase the threshold error at which to refine until fewer than 300 triangles are marked for refinement. Refining a maximum number of triangles helps to stop spurious spreading of the refinement zone and means we obtain a more focused mesh, with less need for derefinement. We chose, after numerical experiments, 300 triangles as an upper refinement limit for the semilinear test case, however a different maximum might be more appropriate for other problems.

We use the popular red/green refinement strategy, as described by Bank *et al.* in [10]. As shown in Figure 5-1 red refinement splits a triangle into four by subdividing the edges of the triangle into two. An extension of this technique would be to also use blue refinement which refines pairs of triangles in such a way that the mesh is quickly oriented to fit any special features of the problem. This has been found to be particularly useful for problems with internal or boundary layers in [46]. However we have not implemented it here.

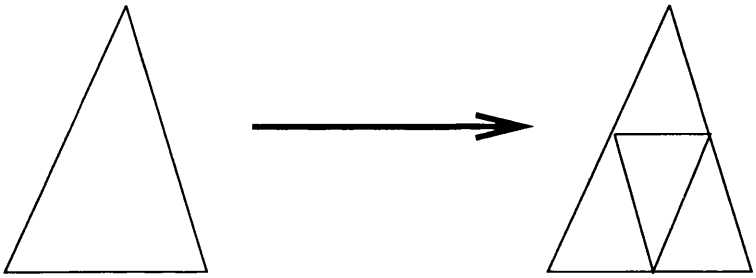


Figure 5-1: Red Refinement. The triangle on the left is called the parent triangle of the four new triangles introduced by the refinement.

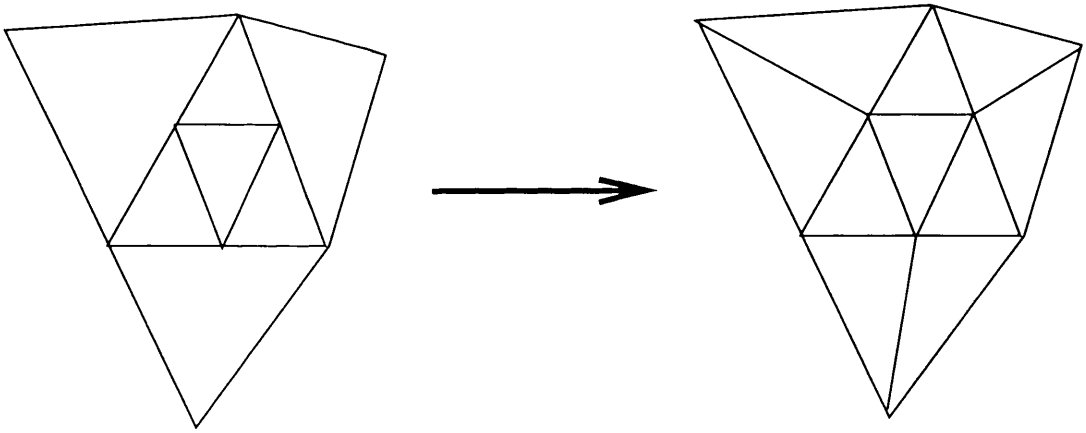


Figure 5-2: Green Refinement, also called green closure.

Derefinement is achieved by removing one level of refinement from the flagged triangle, i.e. the triangle is removed along with other triangles sharing the same parent element from its last refinement, thus we are just left with the previous parent element.

When all the red refinement and the derefinement has been completed we are often left with a non-conforming triangulation (one with hanging mesh points) to prevent

this we then perform green refinement where needed, shown in Figure 5-2. Since green refinement can reduce the size of the minimum angle within the triangulation we do not directly refine a triangle that has been produced in this way, but rather we remove the new edge introduced by the previous green refinement and refine its parent triangle directly using red refinement.

Derefinement is not always possible, for example we cannot remove a triangle formed by green refinement if it would leave a non-conforming triangulation.

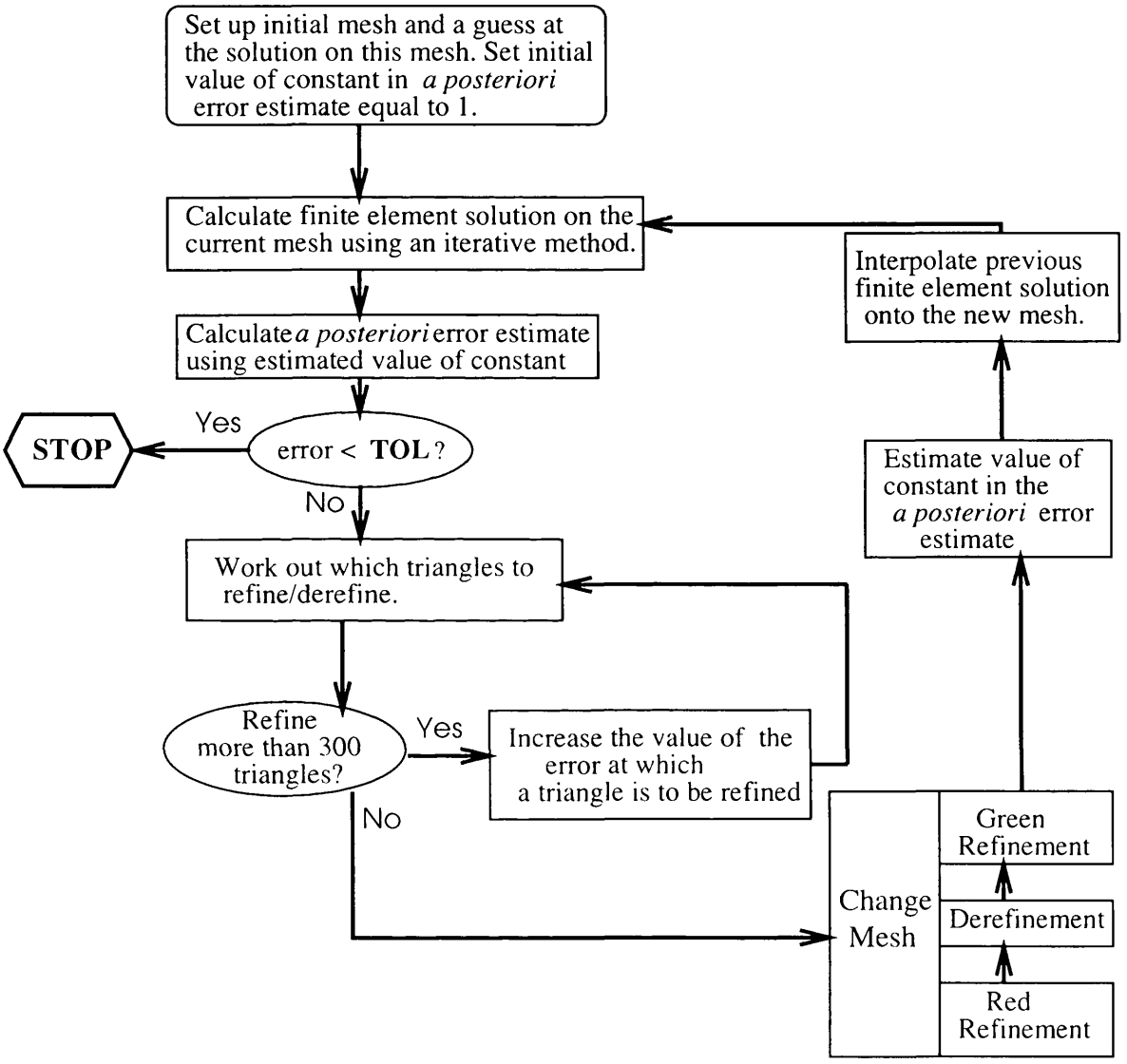


Figure 5-3: The steps involved in calculating an accurate finite element solution using our adaptive refinement procedure.

Remark 5.3.1 *The contribution the difference between the normal derivative of the*

solution over an edge of a triangle makes to the a posteriori error estimate is split equally between the two triangles forming the edge.

In one dimension we are able to implement a slightly different refinement strategy based on [31]. Here we refine an interval ('a one dimensional triangle') into m new intervals, where m depends on the size of the error estimate in the original interval and is typically larger than two. Basically, assuming we have an a posteriori error estimate of the form

$$\|u_0 - u_h\| \leq C \sum_{I_k} \|h_k E(u_h)\|_{I_k}$$

where: $\{I_k\}$ is the set of N_I intervals which we have subdivided our domain into and h_k is the length of the interval I_k, $E(\cdot)$ is some function defining the a posteriori error estimate, $\|\cdot\|$ is a norm on the whole domain and $\|\cdot\|_{I_k}$ is the norm on the interval I_k. We assume the constant C has been estimated. Then, for an interval I_k, if

$$C\|h_k E(u_h)\|_{I_k} > \frac{TOL}{N_I}$$

for some tolerance TOL, then we seek a $\tilde{h}_k := h_k/m_k$, where m_k is an integer such that

$$C\|\tilde{h}_k E(u_h)\|_{I_k} \cong \frac{TOL}{N_I}.$$

The interval I_k is then divided into m_k equal intervals, each with length equal to $\tilde{h}_k$.

We find that decreasing the tolerance slowly at each refinement, rather than starting off using the required tolerance, helps stop the spread of the refinement zone unnecessarily and avoids the need for derefinement.

The constant, C, in the *a posteriori* error estimate is estimated numerically by estimating the norm difference between the finite element solutions at each level of the iteration and comparing it with the values of the *a posteriori* error estimate.

To do this, we assume we have an estimate of the form:

$$\|u_0 - u_h\| \leq C\|\mathrm{E}(u_h)\|.$$

Then, if u_h^k is the finite element solution on the kth mesh, using the triangle inequality we have

$$\|u_h^{k+1} - u_h^k\| \leq C\left\{\|\mathrm{E}(u_h^{k+1})\| + \|\mathrm{E}(u_h^k)\|\right\}. \tag{5.3.57}$$

Since $\left\{\|\mathrm{E}(u_h^{k+1})\| + \|\mathrm{E}(u_h^k)\|\right\}$ is known from our computations of the *a posteriori* error estimate and we may estimate $\|u_h^{k+1} - u_h^k\|$, we can compute a lower bound for C on the $(k+1)$th mesh. This we take as an estimate of C.

The code used to implement this adaptive strategy is discussed in Section 6.1.

5.4 Test Problems for the Adaptive Procedure

5.4.1 The Layer Problem

It is well known that the solution of the problem

$$-\lambda^2 \Delta u + 2\delta^2 \sinh(u) = d \tag{5.4.58}$$

with discontinuous d, exhibits layer behaviour (that is a region of fast variation in the solution) at the junctions of discontinuity in d as $\lambda \to 0+$. Using standard singular perturbation theory a number of authors have calculated the width of the layer at these junctions, see for example [56] for results in one dimension and Section 4.5 of [50] (and the references therein) for results in two dimensions. We test the adaptive procedure introduced in Section 5.3 by trying to capture the width of the layer for a problem of the form (5.4.58).

We consider the following layer problem:

$$-\lambda^2 \Delta u + 2\delta^2 \sinh(u) = d, \qquad \text{in } \Omega = [0,1] \times [0,1], \tag{5.4.59}$$

$$u = \sinh^{-1}\left(\frac{1}{2\delta^2}\right), \qquad \text{on } \partial\Omega_1, \tag{5.4.60}$$

$$u = \sinh^{-1}\left(\frac{-1}{2\delta^2}\right), \qquad \text{on } \partial\Omega_2. \tag{5.4.61}$$

$$\frac{\partial u}{\partial n} = 0, \qquad \text{on } \partial\Omega \setminus \partial\Omega_1 \cup \partial\Omega_2. \tag{5.4.62}$$

where $\partial\Omega_1$ is the set $\{(x,y) : x = 0, y \in [0, \frac{1}{2})\}$, $\partial\Omega_2$ is the set $\{(x,y) : x = 1, y \in [0,1]\}$ and d is the discontinuous function:

$$d(x,y) = \begin{cases} +1 & \text{for } \sqrt{x^2+y^2} < \frac{1}{2} \\ -1 & \text{for } \sqrt{x^2+y^2} > \frac{1}{2} \end{cases}. \tag{5.4.63}$$

For δ^2 we take the value 1×10^{-7}, denote the junction $\sqrt{x^2+y^2} = \frac{1}{2}$ by Γ. With

such parameter values (5.4.58) corresponds to a PN diode (see for example [50]) with geometry represented in Figure 5-4.

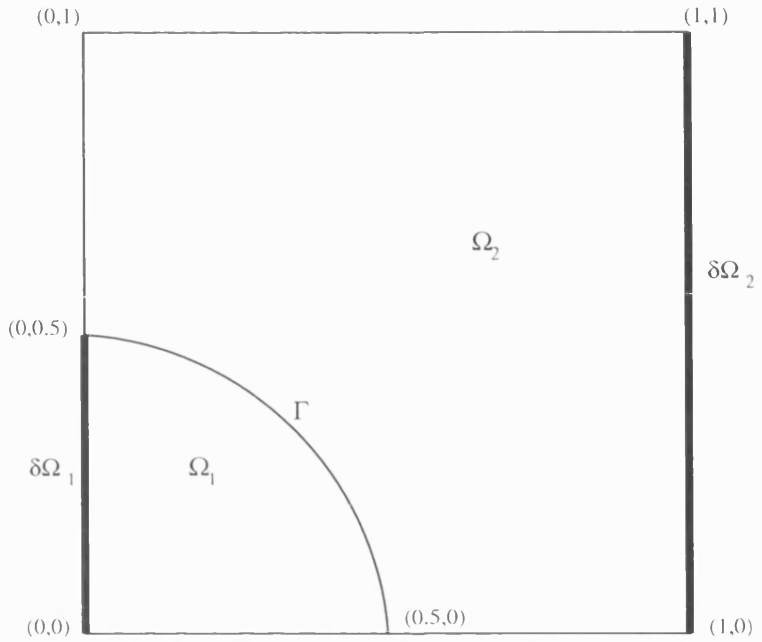


Figure 5-4: The profile of the diode for the test problem. The doping profile takes the value +1 in Ω_1 and -1 in Ω_2. $\Gamma = \left\{(x,y) : (x^2+y^2)^{\frac{1}{2}} = \frac{1}{2}\right\}$ represents the junction between Ω_1 and Ω_2, the layer occurs in a region around Γ. The thick lines represent the Dirichlet boundaries.

The problem (5.4.59)-(5.4.62) with d given by (5.4.63) fits into the framework discussed in [49] with zero applied voltage. In Section 4B of [49] asymptotic analysis is used to compute the width of the layer in u at the junction Γ. It is found that for u satisfying (5.4.59)-(5.4.62) the width of the layer at Γ is of order

$$\lambda \log(\lambda) \tag{5.4.64}$$

as $\lambda \to 0+$.

We try to capture this behaviour numerically using adaptive techniques. In particular as a severe test of adaptivity we try to capture numerically the behaviour, (5.4.64), as $\lambda \to 0+$. One difficulty with this problem is that it is not well defined where the layer begins and ends. We overcome this with our *a priori* knowledge of the solutions to equations of the form (5.4.59)-(5.4.62): from previous numerical experiments it is found that "away" from the layer u essentially has the same value as one of the boundary

conditions (5.4.60), (5.4.61). Thus we say (x, y) is in the layer if

$$\sinh\left(\frac{-1}{2\delta^2}\right) + \epsilon < u(x,y) < \sinh\left(\frac{1}{2\delta^2}\right) - \epsilon \tag{5.4.65}$$

for some small number ϵ. We take $\epsilon = 0.03$ for this experiment.

In our adaptive process we seek a finite element solution, u_h, to (5.4.59)-(5.4.62) satisfying

$$\|u_0 - u_h\|_0 < \text{ TOL},$$

where u_0 is the weak solution and the tolerance, TOL, is set at 5×10^{-3}. We do this by seeking a finite element solution satisfying the L_2 *a posteriori* error estimate

$$\mathrm{E}(u_h) < 5 \times 10^{-3}, \tag{5.4.66}$$

where

$$\mathrm{E}(u_h) = C\left[\left\{\sum_{\mathrm{T}_k \in \mathcal{T}_h} h_k^{2\alpha} A_k^2\right\}^{\frac{1}{2}} + B^2\right]. \tag{5.4.67}$$

In the above B is the H^1 *a posteriori* error estimate and A_k is an estimate of the contribution the triangle $\mathrm{T}_k \in \mathcal{T}_h$ makes to the total H^1 *a posteriori* error estimate. The constant C is the estimate of constant appearing in the *a posteriori* error estimate (estimated in the way described in Section 5.3). As in Section 5.2 define B by:

$$B := \left[\lambda^2 \left\{\sum_{\tau \in \mathcal{E}_h} h_\tau^2 \left[\frac{\partial u_h}{\partial n}\right]_\tau^2\right\}^{\frac{1}{2}} + \left\{\sum_{\mathrm{T}_k \in \mathcal{T}_h} \|h_k f(u_h)\|^2_{\mathrm{L}_2(\mathrm{T}_k)}\right\}^{\frac{1}{2}}\right]$$

and as in (5.2.41), for each $\mathrm{T}_k \in \mathcal{T}_h$:

$$A_k := \left[\lambda^2 \left\{\sum_{\tau \in \mathcal{E}(\mathrm{T}_k)} (h_\tau)^2 \left[\frac{\partial u_h}{\partial n}\right]_\tau^2\right\}^{\frac{1}{2}} + \|h_k f(u_h)\|_{\mathrm{L}_2(\mathrm{T}_k)}\right].$$

Since the estimate (5.4.67) is a combination of the L_2 *a posteriori* error estimate and the H^1 *a posteriori* error estimate, C in (5.4.67) can be seen to be a bound on the constant appearing in the L_2 estimate multiplied by the square of the constant appearing in the H^1 estimate. Due to the nature of the points at which the Dirichlet and Neumann

conditions meet, α in (5.4.67) is taken as 0.5 - see Chapter 4 and [35] for more details.

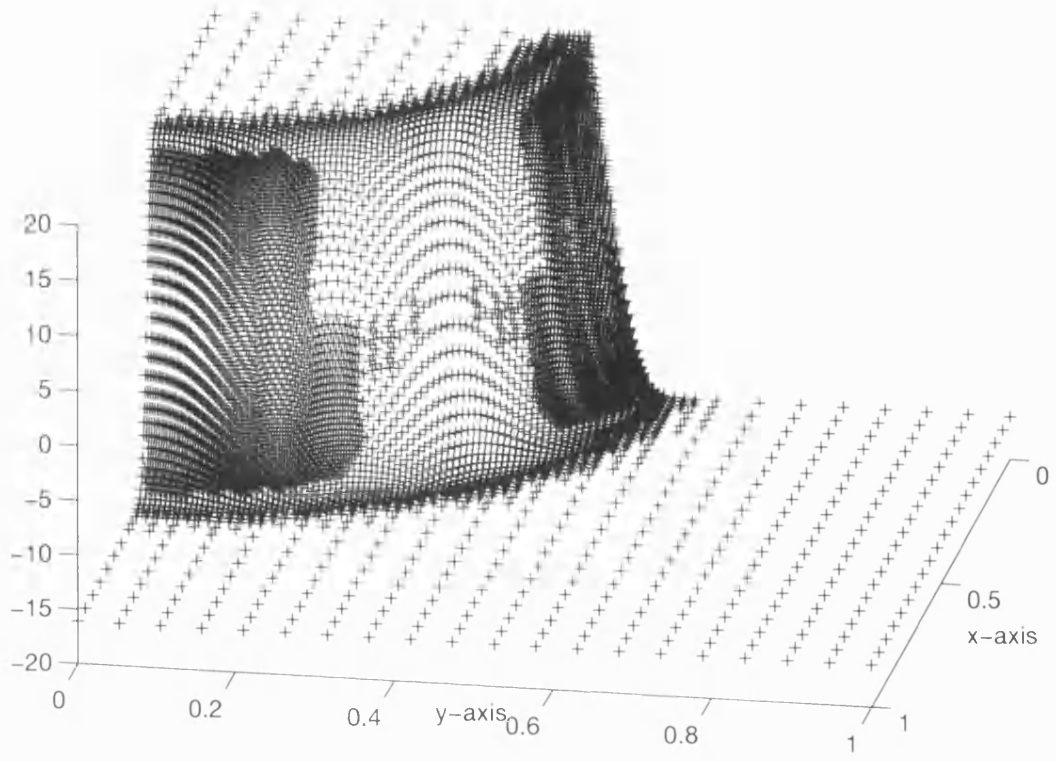


Figure 5-5: The finite element solution to the semilinear problem when $\lambda = 1 \times 10^{-4}$, produced after 18 adaptive refinements. The mesh contains 5947 mesh points and 11761 triangles. The initial mesh was a regular 20×20 mesh.

Using the adaptive scheme outlined in Section 5.3 and the *a posteriori* error estimate from Section 5.2 we are able to solve the finite element system (5.4.59)-(5.4.62), the results are contained in Tables 5.1 and 5.2. A typical picture of the finite element solution when $\lambda = 1 \times 10^{-4}$ is presented in Figure 5-5 and a typical mesh for the same value of λ is shown in Figure 5-6. We note that it is clear from our pictures that our refinement strategy has concentrated the mesh points in the area around the junction Γ where we expect there to be a large change in the true solution.

Table 5.1 shows the finite element solution has a layer of width of order approximately λ as $\lambda \to 0+$, this compares well with the order of width of the layer, (5.4.64), predicted in [49].

It is interesting to look at the value of the estimated constant in the *a posteriori*

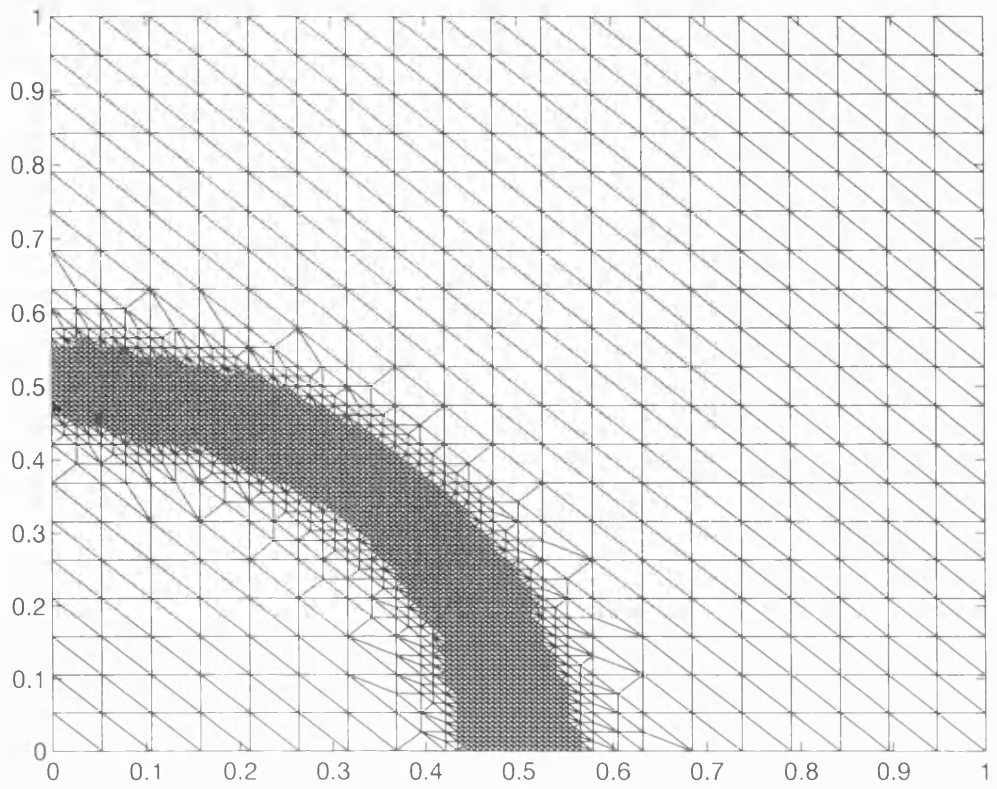


Figure 5-6: The mesh for the semilinear problem when $\lambda = 1 \times 10^{-4}$. This mesh is produced after 9 adaptive refinements and contains 2763 mesh points and 5413 triangles. The initial mesh was a regular mesh of size 20×20.

λ^2	Size of initial mesh	Number of refinements	Final number of mesh points	Width of layer	Order of λ in width
1×10^{-2}	10×10	23	5013	0.9722	—
1×10^{-3}	10×10	14	3247	0.5277	0.5307
5×10^{-4}	10×10	11	2230	0.3611	1.0946
1×10^{-4}	10×10	15	2963	0.1527	1.0695
5×10^{-5}	10×10	12	3894	0.1111	0.9177
1×10^{-5}	20×20	16	6453	0.0526	0.8679
5×10^{-6}	20×20	12	3667	0.0382	0.9230
1×10^{-6}	30×30	10	4166	0.0193	0.8484

Table 5.1: shows how the numerically computed width of the layer depends on λ as $\lambda \to 0+$. The theory predicts that the width is of order $\lambda \log(\lambda)$ as $\lambda \to 0+$. These results are computed using the L_2 *a posteriori* error estimates and a tolerance of 5×10^{-3}.

λ^2	Estimate of the constant C appearing in (5.4.67)	Order of λ in the estimate of the constant
1×10^{-2}	0.83	—
1×10^{-3}	3.65	-1.28
5×10^{-4}	6.25	-1.55
1×10^{-4}	12.13	-0.82
5×10^{-5}	16.94	-0.96
1×10^{-5}	68.33	-1.73
5×10^{-6}	98.73	-1.06
1×10^{-6}	328.92	-1.49

Table 5.2: shows how the estimate of the constant computed numerically on a mesh adapted from a regular 10×10 mesh, depends on λ. The theory predicts an order of $(\lambda)^{-6}$ as $\lambda \to 0+$, but this is not seen in practice. The value of the constant given is an average of the computed values as we near the required tolerance.

error estimate and compare it with the value one would predict from the theory:

From Table 5.2 we see that the order of λ in the estimate of the constant in the L_2 *a posteriori* error estimate (5.4.67) grows with order between $O(\lambda^{-1})$ and $O(\lambda^{-2})$ as $\lambda \to 0+$. Since the constant, C, in (5.4.67) is a bound on the constant appearing in the L_2 estimate multiplied by the square of the constant appearing in the H^1 estimate and each of these constants is of order $O(\Lambda^0)$, it appears that the constant may grow as fast as:

$$O\left(\Lambda^0\right)^3.$$

Since the leading term in the Fréchet derivative of (5.4.58) is: $-\lambda^2 \Delta u$, one may expect

$$\Lambda^0 = O(\lambda^{-2})$$

and thus the constant in the *a posteriori* error estimate may be estimated to be of order λ^{-6} if standard analysis is applied. The observed order of λ^{-2} shows that standard estimates may be very pessimistic.

The reason for the difference in this case may be partly explained by the following heuristic argument:

The Fréchet derivative associated with our problem is:

$$(\mathrm{F}'(u_0)v, w) = (\lambda^2 \nabla v, \nabla w) + (2\delta^2 \cosh(u_0)v, w).$$

We aim to show that the effective numerical inverse of the discretised Fréchet derivative (i.e. the constant Λ^0 above) does not blow up with order λ^{-2} as the analytic theory suggests. To see this consider discretising the Fréchet derivative using piecewise linear finite elements on a uniform mesh with mesh size h (and mass lumping the zero order term), doing this we obtain a matrix of the following form :

$$\lambda^2 K + 2\delta^2 h^2 \begin{bmatrix} \cosh(u_0(0)) & & & & \\ & \cosh(u_0(1)) & & & \\ & & \ddots & & \\ & & & \cosh(u_0(n-1)) & \\ & & & & \cosh(u_0(n)) \end{bmatrix}, \tag{5.4.68}$$

where $u_0(i)$ represents the value of the weak solution at mesh point i and K is the finite element stiffness matrix corresponding to the Laplacian.

As we have observed, for a mesh point i, $u_0(i)$ is either close to the Dirichlet boundary conditions ($u_0(i) \simeq \sinh^{-1}(\pm 1/2\delta^2)$) or i is in the layer. For a mesh point i not in the layer, since δ is small:

$$\begin{aligned} \cosh(u_0(i)) &\simeq \cosh\left(\sinh^{-1}\left(\frac{\pm 1}{2\delta^2}\right)\right) \\ &\simeq \cosh\left(\cosh^{-1}\left(\frac{1}{2\delta^2}\right)\right) \\ &= \frac{1}{2\delta^2}. \end{aligned}$$

Therefore, for a mesh point i not in the layer the ith row of the matrix (5.4.68) is essentially the ith row of the matrix $\lambda^2 K + h^2 I$, which is dominated by the diagonal matrix $h^2 I$ when λ is small. However, for a mesh point i in the layer $u_0(i)$ is typically small compared to the boundary conditions and

$$\cosh(u_0(i)) \simeq 1.$$

The ith row of (5.4.68) is then essentially the ith row of $\lambda^2 K + 2\delta^2 h^2 I$. Assuming $\delta < \lambda$ this is of order $O(\lambda^2)$ as $\lambda \to 0$, but this only happens in the layer.

We conclude that this heuristic argument suggests that the finite element approximation of $\|(\mathrm{F}'(u_0))^{-1}\|$ should vary in size between $O(h^{-2})$ and $O(\lambda^{-2})$. $\|(\mathrm{F}'(u_0))^{-1}\|$

will only approach $O(\lambda^{-2})$ when we have refined into the layer significantly and the $O(\lambda^{-2})$ terms dominate the $O(h^{-2})$ terms. This may explain why we are only seeing an order of λ^{-2}, in the estimate of the constant, rather than the order λ^{-6} the theory predicts.

5.4.2 Efficiency of the Adaptive Method

It is also interesting to look at the efficiency of an adaptive method, i.e. how close the *a posteriori* error estimate gets to the real error in the finite element solution as we refine the mesh. We make the following definition of efficiency of an adaptive method:

Definition 5.4.1 *Let u_0 be the weak solution of a given problem and let u_h^k be the corresponding finite element solution on the kth mesh. Then for a given norm, $\|\cdot\|$, define*

$$\textit{efficiency of the method} := \lim_{k\to\infty} \frac{\textit{true value of } \|u_0 - u_h^k\|}{\textit{upper bound on } \|u_0 - u_h^k\|}.$$

Ideally the efficiency should be close to 1, showing that the method is correctly estimating the error in the solution as the scheme progresses.

The efficiency of *a posteriori* error estimators is also considered in [2], where the efficiency of Bank and Weiser's error estimates ([11]) are theoretically tested for degree p finite element approximations on quadrilateral meshes. It is shown there that the error estimators are asymptotically exact (as the mesh diameter tends to zero) for regular problems, providing that the degree of approximation is of odd order and the elements are rectangular.

We test the efficiency of our adaptive method on the following semilinear problem:

$$-\Delta u + \frac{2(1.9)^2}{(1-(1.9x-0.95)^2)^2}\tanh(u) = 0, \qquad \text{in } \Omega = [0,1]\times[0,1], \quad (5.4.69)$$

$$u(x,y) = -\tanh^{-1}(0.95). \qquad \text{on } \partial\Omega_1. \quad (5.4.70)$$

$$u(x,y) = \tanh^{-1}(0.95). \qquad \text{on } \partial\Omega_2, \quad (5.4.71)$$

$$\frac{\partial u}{\partial n} = 0. \qquad \text{on } \partial\Omega \setminus \partial\Omega_1 \cup \partial\Omega_2. \quad (5.4.72)$$

where $\partial\Omega_1 = \{0\}\times[0,1]$ and $\partial\Omega_2 = \{1\}\times[0,1]$.

The problem (5.4.69)-(5.4.70) has solution

$$u(x,y) = \tanh^{-1}(1.9x - 0.95). \quad (5.4.73)$$

Refinement level	L_2 *a posteriori* error estimate	Efficiency at this refinement level
1	1.28×10^{-2}	2.92
2	2.67×10^{-3}	1.13
3	3.29×10^{-3}	3.84
4	3.24×10^{-3}	1.01
5	5.01×10^{-3}	4.28
6	5.42×10^{-4}	3.13
7	1.27×10^{-3}	2.23
8	8.53×10^{-4}	2.61
9	6.29×10^{-4}	2.85
10	3.21×10^{-4}	4.94
11	3.24×10^{-4}	4.30
12	5.84×10^{-4}	4.23
13	4.20×10^{-4}	2.61
14	1.52×10^{-4}	1.29

Table 5.3: shows the efficiency of the adaptive method as we refine the mesh. The values of the *a posteriori* error estimates shown include the estimated value of the constant at each refinement.

which has slight boundary layers at $\partial\Omega_1$ and $\partial\Omega_2$.

We use our adaptive method to compute a finite element solution to the problem and see how the efficiency of the method behaves as we refine the mesh. We use the L_2 norm in our test of efficiency as it is reasonably easy to approximate the true error $\|u_0 - u_h\|_0$ using (B.0.1), where u_0 is given by (5.4.73). The upper bound on $\|u_0 - u_h\|_0$ consists of the *a posteriori* error estimate (5.2.26) with the constant estimated as described in Section 5.3.

The results, presented in Table 5.3, show that although the efficiency of the adaptive method applied to (5.4.69)-(5.4.72) does not conclusively tend towards 1, the *a posteriori* error estimate continually over estimates the error in the finite element solution. The variation in the efficiency value (and in the size of the L_2 *a posteriori* error estimate) is due to variation in the estimated value of the constant in the error estimate. One way to smooth this variation might be to take an average of the recently estimated values of the constant as we refine the mesh, but we should probably also make sure any average has a bias towards the last estimated value.

Chapter 6

Efficient Adaptive Numerical Models of Typical Semiconductor Devices

In this chapter we test the methods introduced in the two previous chapters on two very different semiconductor problems: A PIN diode in its off-state and a MOSFET diode with varying applied voltages.

In Section 6.2 we consider the PIN diode problem. With zero applied voltage the model reduces to a single semilinear equation with two small parameters, λ and δ. The limiting forms of the solution as $\lambda \to 0$ and as $\delta \to 0$ are known from asymptotic analysis. We use our *a posteriori* error estimates to produce accurate finite element solutions to the semilinear equation. We show the effectiveness of the refinement procedure by demonstrating that as $\lambda \to 0$ or $\delta \to 0$ the numerical solutions have the right asymptotic behaviour. These initial results are obtained by solving the full nonlinear problem on each of the fine meshes.

We also test our defect correction method on this problem, but use adaptively determined meshes, rather than the *a priori* determined meshes of the theory in Chapter 4. The defect correction method solves a nonlinear problem on the coarsest mesh and one linear problem on each of the finer grids. We show that this adaptive defect correction method is competitive with the method that solves a nonlinear problem on each of the meshes. It is shown that the method produces solutions of the correct form, provided the initial mesh is sufficiently fine and the meshes are refined cautiously.

In Section 6.3 we consider the MOSFET diode with four contacts. Using a series of simplifying assumptions we reduce the problem to a system which is easier to solve. This system is solved for a variety of applied voltages and the calculated electron concentrations compared to the known behaviour of the electrons in the MOSFET.

First we provide some details of the program used to find the finite element solutions:

6.1 The Finite Element Code

The adaptive finite element code used in this chapter and for the PN diode experiments in Chapter 5 combines and extends two research codes: PETSc [63] and FEMLAB [38].

PETSc [http://www.mcs.anl.gov/petsc/petsc.html] is intended for use in large scale applications and has an extensive range of tools for the numerical solution of partial differential equations. The code can be used on machines set up in parallel or serial, though we only use it on a single machine. PETSc is written in C and uses the MPI standard for message passing. We have extended the code to include unstructured grids of triangles and the discretisation of the operators appearing in the semiconductor equations. In our code PETSc is used to discretise the finite element problems and solve the resulting systems.

FEMLAB [http://www.math.chalmers.se/Research/Femlab/index.html] is a less powerful and much slower Fortran code designed for solving convection-diffusion problems but has a very good adaptive refinement procedure. We use the grid structure and refinement code from FEMLAB, adding our own *a posteriori* error estimates and refinement criteria. The refinement strategy in the extended code is the procedure described in Section 5.3.

The code operates in the following way:

(1) Set up the initial grid and guesses at the solutions to the equations. (FEMLAB)

(2) Set up and discretise the operators, adding in the boundary conditions. (PETSc extension)

(3) Solve the resulting systems. (PETSc)

(4) Calculate the *a posteriori* error estimate and which triangles to refine. (FEMLAB extension)

(5) Refine the grid if the *a posteriori* error estimate is greater than the tolerance and return to (2). Otherwise exit the code. (FEMLAB)

The adaptive finite element code is used in the next two sections.

6.2 The PIN Diode in Thermal Equilibrium

The PIN diode is a semiconductor device with an n and p region separated by an intrinsic (or i) region. An i region has a very low concentration of ionized impurities (characterised by a zero or approximately zero doping profile in the region). The device behaves like a PN diode but has some additional features. In this section two numerical methods for solving the finite element system associated with the PIN diode in thermal equilibrium are tested against each other and the results compared to the asymptotic analysis contained in Section 4.4 of [51].

6.2.1 The PIN Diode Equations in Thermal Equilibrium

The two dimensional semiconductor device equations discussed in Section 1.3 with a zero applied voltage across the device (thermal equilibrium) reduce to the problem:

$$-\lambda^2\Delta\psi + 2\delta^2\sinh(\psi) - d = 0, \qquad \text{in } \Omega \subset \mathbb{R}^2, \tag{6.2.1}$$

$$\psi = \sinh^{-1}\left(\frac{d|_{\partial\Omega_D}}{2\delta^2}\right), \qquad \text{on } \partial\Omega_D, \tag{6.2.2}$$

$$\frac{\partial\psi}{\partial n} = 0, \qquad \text{on } \partial\Omega_N. \tag{6.2.3}$$

The Dirichlet boundary, $\partial\Omega_D$, and the Neumann boundary, $\partial\Omega_N$, are two disjoint regions whose union is the whole of the boundary of Ω. In (6.2.1), d is the doping profile and for the PIN diode considered in this section will take the form:

$$d = \begin{cases} +1 & \text{in } \Omega_+ \\ 0 & \text{in } \Omega_0 \\ -1 & \text{in } \Omega_- \end{cases}. \tag{6.2.4}$$

Ω_+, Ω_0 and Ω_- are disjoint, simply connected sub-domains of Ω with

$$\overline{\Omega}_+ \cup \overline{\Omega}_0 \cup \overline{\Omega}_- = \overline{\Omega}, \quad \overline{\Omega}_+ \cap \overline{\Omega}_- = \emptyset.$$

In this section Ω is taken to be the unit square and

$$\begin{aligned}
\Omega_+ &:= \{(x,y) \in \mathbb{R}^2 : 0.75 < x \leq 1,\ 0 \leq y \leq 1\},\\
\Omega_- &:= \{(x,y) \in \mathbb{R}^2 : 0 \leq x < 0.25,\ 0 \leq y \leq 0.5 \text{ or } \sqrt{x^2 + (y-0.5)^2} < 0.25\},\\
\Omega_0 &:= \Omega \backslash (\overline{\Omega}_+ \cup \overline{\Omega}_-).
\end{aligned}$$

The Dirichlet boundary of the domain, $\partial\Omega_D$, is split into two parts: Γ_+ and Γ_-:

$$\begin{aligned}
\Gamma_+ &:= \{(1,y) : 0 \leq y \leq 1\},\\
\Gamma_- &:= \{(0,y) : 0 \leq y < 0.5\}.
\end{aligned}$$

For the doping profile defined by (6.2.4) the Dirichlet boundary conditions (6.2.2) are

$$\psi = \sinh^{-1}\left(\frac{1}{2\delta^2}\right) \qquad \text{on } \Gamma_+, \tag{6.2.5}$$

$$\psi = \sinh^{-1}\left(\frac{-1}{2\delta^2}\right) \qquad \text{on } \Gamma_-. \tag{6.2.6}$$

The domain, sub-domains and Dirichlet boundaries are shown in Figure 6-1.

6.2.2 Asymptotic Analysis for the PIN Diode

Approximations to the solution of the PIN diode problem (6.2.1)-(6.2.4) are obtained by exploiting the smallness of λ and δ in [51].

In the PN diode case, as discussed in Section 5.4, the limits $\lambda \to 0$ and $\delta \to 0$ commuted in the asymptotic analysis of [49]. This is not true for the PIN diode. Different approximate solutions are obtained as $\lambda \to 0$ and $\delta \to 0$. The following asymptotic results are obtained in Section 4.4 of [51].

When $\lambda < \delta$ and letting $\lambda \to 0$ the asymptotic solution to (6.2.1)-(6.2.4) is calculated to be:

$$\psi = \begin{cases} \sinh^{-1}\left(\frac{1}{2\delta^2}\right) & \text{in } \Omega_+ \\ 0 & \text{in } \Omega_0 \\ \sinh^{-1}\left(\frac{-1}{2\delta^2}\right) & \text{in } \Omega_- \end{cases}. \tag{6.2.7}$$

However, when $\delta < \lambda$ and $\delta \to 0$ the asymptotic analysis gives the following solution

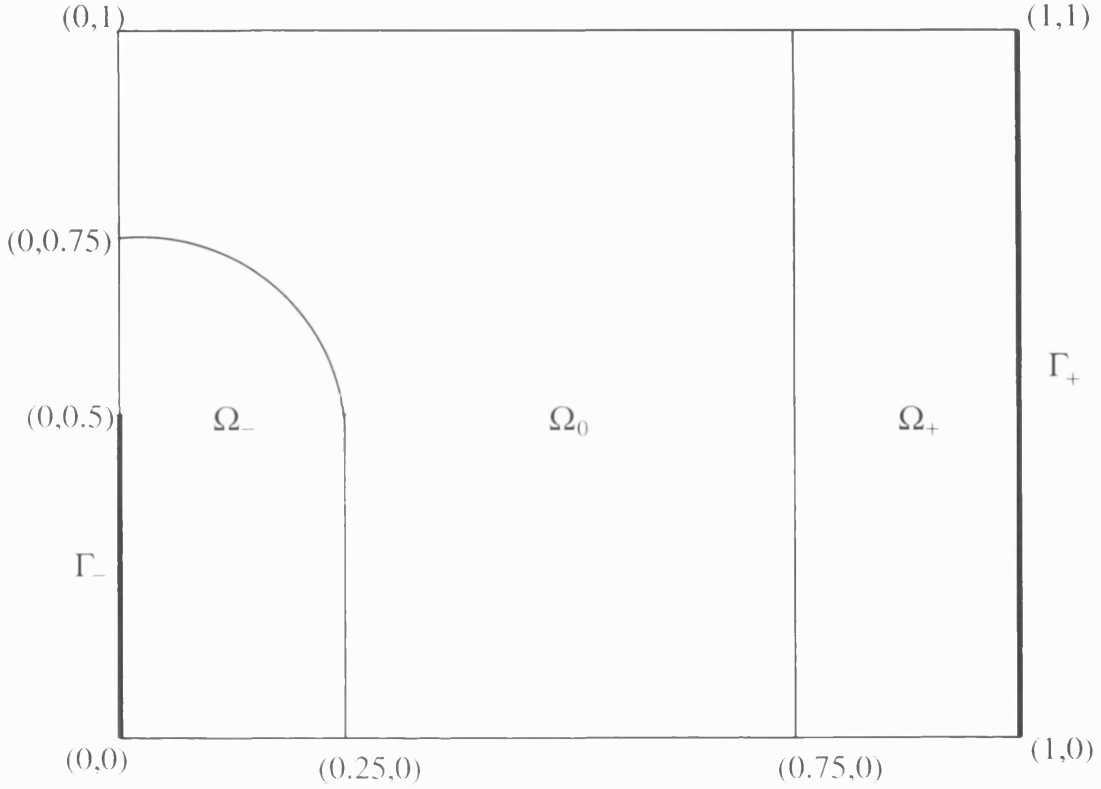


Figure 6-1: Cross section of the PIN diode considered.

to (6.2.1)-(6.2.4):

$$\psi|_{\Omega_+} = \sinh^{-1}\left(\frac{1}{2\delta^2}\right),\ \psi|_{\Omega_-} = \sinh^{-1}\left(\frac{-1}{2\delta^2}\right) \text{ and } \Delta\psi = 0 \text{ in } \Omega_0. \tag{6.2.8}$$

6.2.3 Numerical Results for the PIN Diode Problem

This section compares the efficiency of the two numerical methods introduced in Chapters 5 and 4. The ability of each of the methods to capture the asymptotic results detailed in Section 6.2.2 is also examined.

The first method is the standard finite element method with adaption where a series of nonlinear finite element problems are solved on a series of refined grids, the grid refinement is determined by using an *a posteriori* error estimate for the problem, as discussed in Chapter 5.

The second method is the defect correction method given in Chapter 4, this solves one nonlinear finite element problem and then a sequence of linear finite element problems on carefully refined grids, the grid refinement is determined by carefully using the same *a posteriori* error estimate.

For convenience the mass lumped version (see (6.2.9) below) of the finite element method is used for both schemes.

The Standard Finite Element Method with Adaption

Given a grid of triangles $\mathcal{T}_h = \{T_k\}$, define $\mathcal{V}_h$ to be the piecewise linear finite element space associated with the grid. The mass lumped finite element method for the PIN diode problem (6.2.1)-(6.2.4) is to seek $\psi_h \in \mathcal{V}_h$ such that:

$$(\mathrm{F}(\psi_h), v_h) := (\lambda^2 \nabla\psi_h, \nabla v_h) + \langle 2\delta^2 \sinh(\psi_h) - d, v_h\rangle = 0, \quad v_h \in \mathcal{V}_h, \tag{6.2.9}$$

where $(\cdot,\cdot)$ is the L_2 inner product and, denoting the mesh points of a triangle T_k by p and its area by $\mathcal{A}(T_k)$, the mass lumped inner product is defined by:

$$\langle v, w\rangle := \sum_{T_k \in \mathcal{T}_h} \frac{\mathcal{A}(T_k)}{3} \sum_{p \in T_k} v(p)w(p).$$

(6.2.9) is equivalent to seeking $\psi_h \in \mathcal{V}_h$ such that

$$\mathrm{F}(\psi_h) = 0 \ \text{ in } \ (\mathcal{V}_h)' \tag{6.2.10}$$

where $(\mathcal{V}_h)'$ is the dual space of $\mathcal{V}_h$.

Define ψ_0 to be the weak solution of (6.2.1)-(6.2.4). The *a posteriori* error estimates associated with (6.2.9) are (see Chapter 5):

$$\begin{aligned}
\|\psi_0 - \psi_h\|_1 &\leq C_1 \left[\lambda^2 \left\{ \sum_{\tau \in \mathcal{E}_h} h_\tau^2 \left[\frac{\partial u_h}{\partial n} \right]_\tau^2 \right\}^{\frac{1}{2}} + \left\{ \sum_{\mathrm{T}_k \in \mathcal{T}_h} \|h_k f(u_h)\|^2_{\mathrm{L}_2(\mathrm{T}_k)} \right\}^{\frac{1}{2}} \right] \\
\|\psi_0 - \psi_h\|_0 &\leq C_2 \left[\left\{ \sum_{\mathrm{T}_k \in \mathcal{T}_h} h_k^{2\alpha} \|u_0 - u_h\|^2_{\mathrm{H}^1(\mathrm{T}_k)} \right\}^{\frac{1}{2}} + \|u_0 - u_h\|_1^2 \right].
\end{aligned} \tag{6.2.11}$$

In (6.2.11) $\mathcal{E}_h = \{\tau\}$ is the set of edges of the triangles in $\mathcal{T}_h$, h_k is the diameter of the triangle $T_k \in \mathcal{T}_h$ and $\alpha \in (0.25, 1]$ is a fixed positive constant depending purely on the domain and boundary conditions. α can be calculated using the results of [35]. With the boundary conditions given in Section 6.2.1 $\alpha = 0.5$.

C_1 and C_2 in the *a posteriori* error estimates depend on λ, but for the purposes of this set of experiments C_1 and C_2 are taken to be equal to one. As discussed in Chapter 5 C_2 may theoretically be of order λ^{-6} (numerical experiments in Section 5.4.1 for the PN diode suggest C_2 is of order λ^{-2}), therefore taking $C_2 = 1$ in our numerical method is equivalent to seeking a finite element solution which has an L_2 error significantly less than the required tolerance. Since we are mainly interested in comparing the performance of the numerical methods at the same values of λ and δ this will not affect the results.

The numerical algorithm for the standard adaptive method is:

(1) Choose an initial finite element mesh and a tolerance.

(2) Seek a finite element solution to (6.2.9) on the current mesh using Newton's method.

(3) If the error in the finite element solution, as measured by the L_2 *a posteriori* error estimate, is less than the tolerance then stop. Otherwise, refine the mesh based on the *a posteriori* error estimate. Return to (2).

As in Section 5.4 refinement is based on the L_2 *a posteriori* error estimate (6.2.11).

A triangle is refined if its error (indicated by the size of the contribution the triangle makes to the total *a posteriori* error estimate) exceeds the average error taken over all the triangles by some factor. To avoid over-refinement a maximum of 500 triangles are refined at each loop of the algorithm.

δ^2	Initial mesh	Final number of mesh points	Final number of triangles	Number of refinement steps
1×10^{-4}	10×10	4155	8205	9
5×10^{-5}	10×10	3157	6212	8
1×10^{-5}	10×10	4866	9624	9
5×10^{-6}	10×10	4823	9540	9
1×10^{-6}	20×20	3222	6324	7
5×10^{-7}	20×20	3172	6225	7
1×10^{-7}	20×20	3286	6454	7
5×10^{-8}	20×20	3442	6763	7
1×10^{-8}	20×20	3904	7686	7

Table 6.1: Results for the standard adaptive method with $\lambda^2 = 1 \times 10^{-4}$ and $\delta \to 0$. The solutions are for a tolerance of 5×10^{-3}.

λ^2	Initial Mesh	Final number of mesh points	Final number of triangles	Number of refinement steps
1×10^{-4}	10×10	4155	8205	9
5×10^{-5}	10×10	3150	6203	9
1×10^{-5}	10×10	5087	10074	12
5×10^{-6}	20×20	6704	13276	11
1×10^{-6}	20×20	4764	9415	13
5×10^{-7}	20×20	4987	9860	10
1×10^{-7}	20×20	6898	13681	11
5×10^{-8}	20×20	6943	13770	11
1×10^{-8}	20×20	3768	7427	10

Table 6.2: Results for the standard adaptive method with $\delta^2 = 1 \times 10^{-4}$ and $\lambda \to 0$. The solutions are for a tolerance of 5×10^{-3}.

Using the numerical algorithm finite element solutions to the PIN diode in thermal equilibrium were computed for a variety of λ and δ. The results are contained in Tables 6.1 and 6.2. Selected finite element solutions are shown in Figures 6-2, 6-3, 6-5 and 6-6.

It is observed and can be seen from Figures 6-2 and 6-3, that as $\delta \to 0$ the solution in the regions Ω_+ and Ω_- are approximately equal to the boundary conditions as predicted.

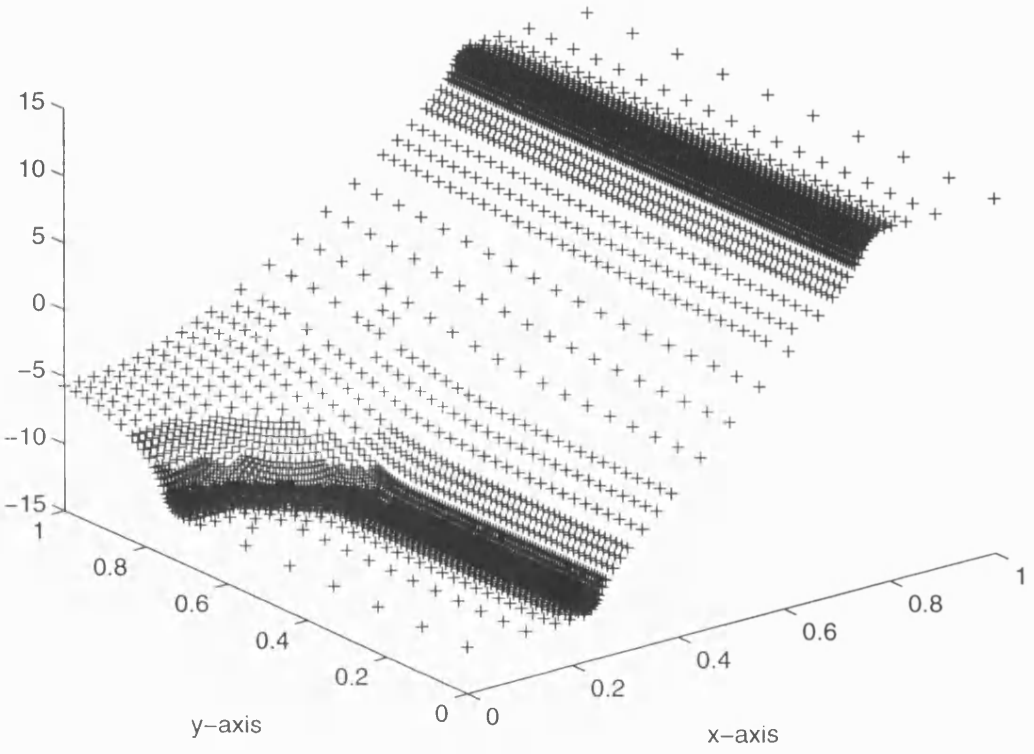


Figure 6-2: The finite element solution to the PIN diode problem when $\lambda^2 = 1 \times 10^{-4}$ and $\delta^2 = 1 \times 10^{-5}$.

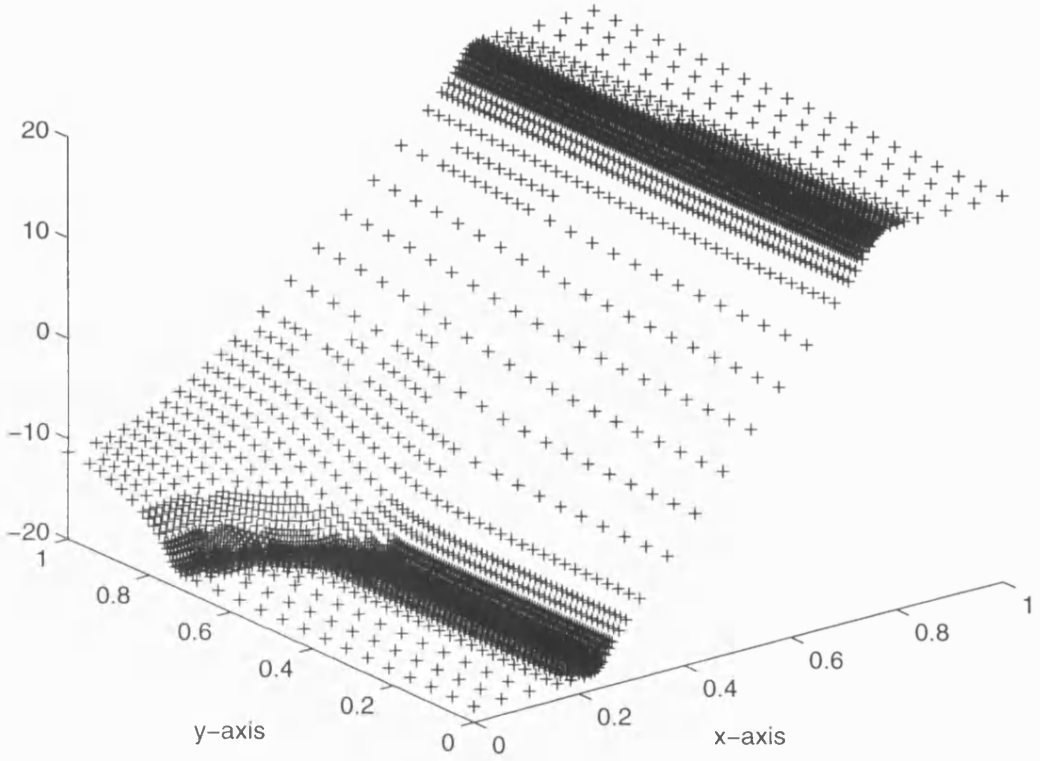


Figure 6-3: The finite element solution to the PIN diode problem when $\lambda^2 = 1 \times 10^{-4}$ and $\delta^2 = 1 \times 10^{-8}$.

In Figure 6-4 the discrete Laplace operator applied to the finite element solution is shown. The maximum value of the Laplacian of the finite element solution is of order 10^{-4} in the interior of Ω_0. Thus as $\delta \to 0$ the adapted finite element solution is clearly of the form predicted by (6.2.8).

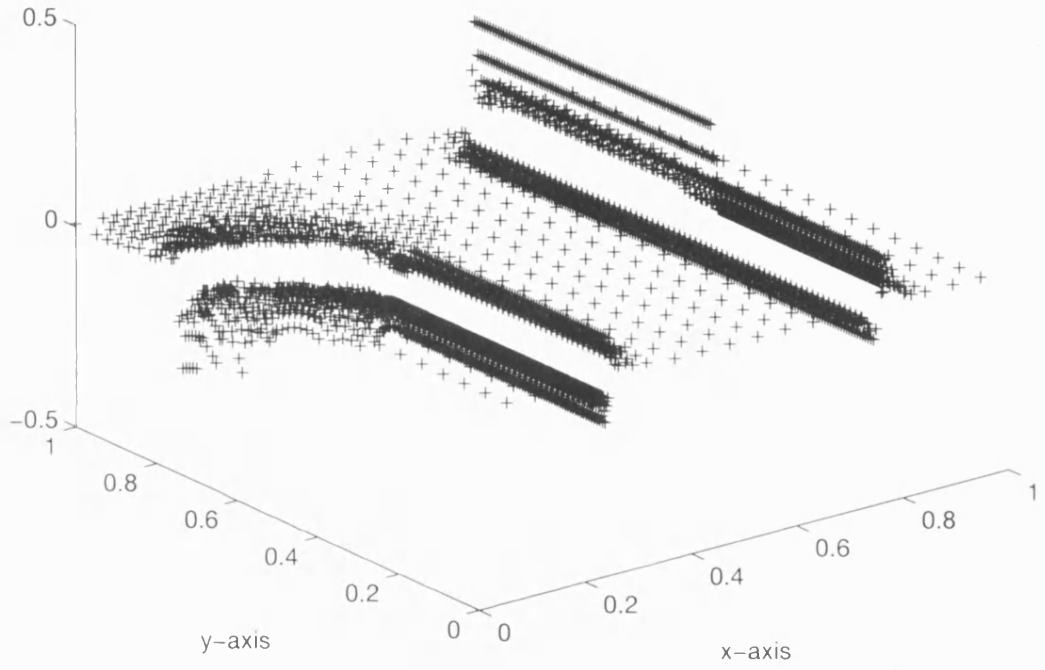


Figure 6-4: The discretised Laplace operator applied to the best finite element solution of the PIN diode problem when $\lambda^2 = 1 \times 10^{-4}$ and $\delta^2 = 1 \times 10^{-8}$.

Figures 6-5 and 6-6 show the finite element solution of the PIN diode problem when $\delta^2 = 1 \times 10^{-4}$ and $\lambda^2 = 1 \times 10^{-5}$ and 1×10^{-8}. It can be seen from the pictures that as $\lambda \to 0$ the finite element solution is of the type predicted by the asymptotic analysis.

The layers in the finite element solution as $\lambda \to 0$ are much deeper than when $\delta \to 0$, this is reflected in the need for a mesh with more mesh points as $\lambda \to 0$. It seems strange that fewer mesh points are needed when $\lambda^2 = 1 \times 10^{-8}$ than when $\lambda^2 = 5 \times 10^{-8}$, but it would appear that this is due to the increased sharpness of the layers and flatness of

the solution in the part of the domain corresponding to Ω_0. [Very few mesh points are needed to capture layers that are almost vertical drops].

As $\lambda \to 0$ we see from (6.2.7) that the solution should become more like a scaled version of the doping profile (scaled according to the boundary conditions), in particular the layers in the finite element solution should occur at approximately the same points in the domain as the jumps in the doping profile. To test this the starting positions of the left and right hand layers were computed. The right hand layer is defined to start when the solution is bounded away from the Dirichlet boundary condition at $x = 1$ by a factor of 0.03 [0.03 is an arbitrary choice, any small number could have been used here]. The left hand layer is defined to start when the finite element solution in the region $\{(x, y) : x \in [0, 1], y \in [0, 0.45)\}$ is bounded away from the Dirichlet condition at $x = 0$ by 0.03. In theory, as $\lambda \to 0$, these should tend towards 0.25 and 0.75 respectively. The results of this test are contained in Table 6.3. The results do indeed conform with the theory and illustrate the sharpness of the finite element solution.

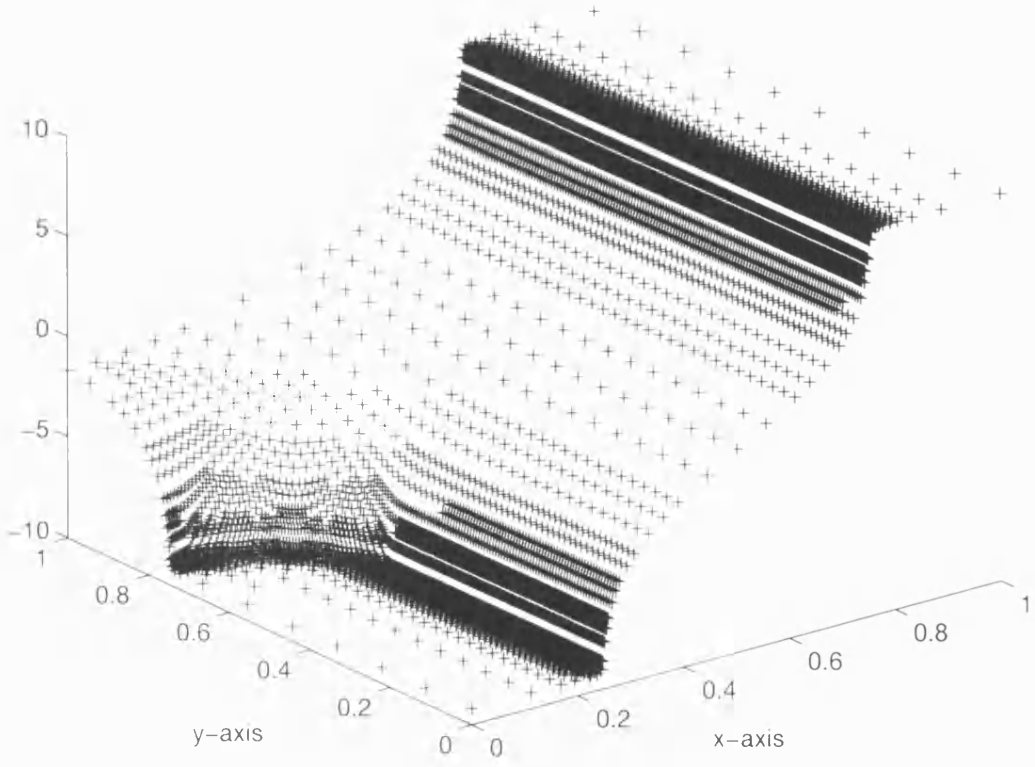


Figure 6-5: The finite element solution to the PIN diode problem when $\delta^2 = 1 \times 10^{-4}$ and $\lambda^2 = 1 \times 10^{-5}$.

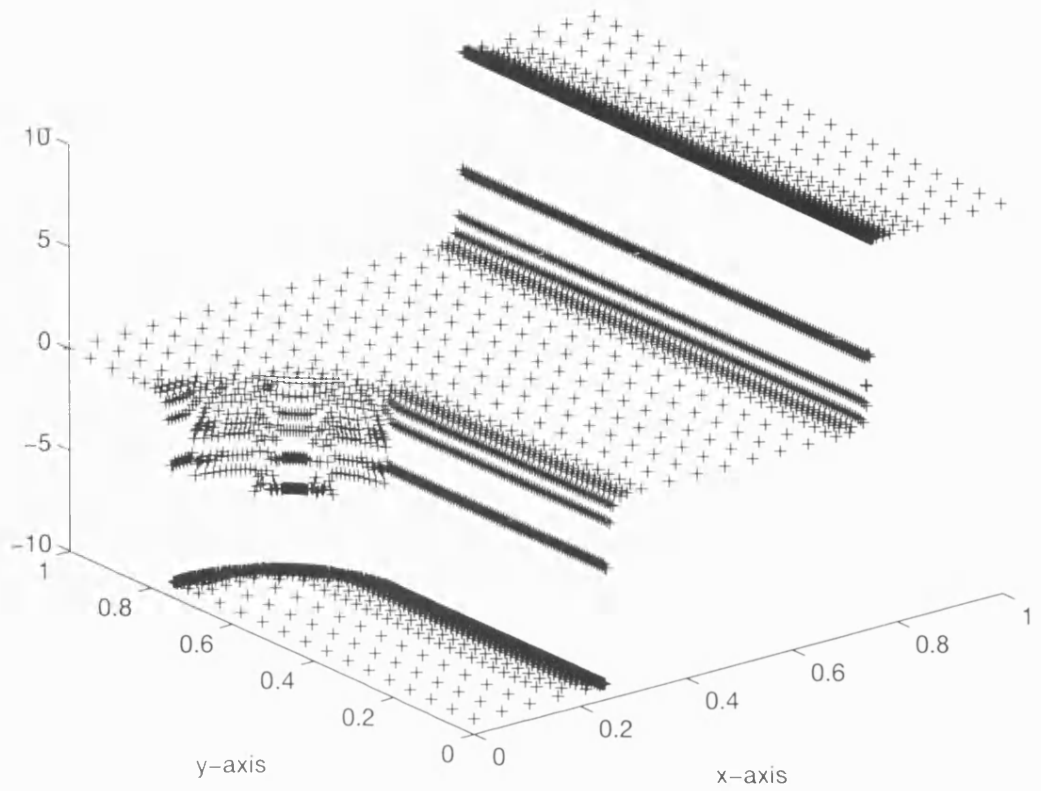


Figure 6-6: The finite element solution to the PIN diode problem when $\delta^2 = 1 \times 10^{-4}$ and $\lambda^2 = 1 \times 10^{-8}$.

λ^2	Position of right hand layer	Position of left hand layer
1×10^{-4}	0.2222	0.7777
5×10^{-5}	0.2222	0.7777
1×10^{-5}	0.2431	0.7604
5×10^{-6}	0.2401	0.7599
1×10^{-6}	0.2467	0.7533
5×10^{-7}	0.2467	0.7533
1×10^{-7}	0.2484	0.7516
5×10^{-8}	0.2484	0.7516
1×10^{-8}	0.2500	0.7500

Table 6.3: The x positions of the layers in the finite element solution as $\lambda \to 0$ and $\delta = 1 \times 10^{-4}$. The layers are defined to start when the finite element solution is bounded away from the boundary conditions.

The Defect Correction Method

In this section the defect correction method, introduced in Chapter 4, will be applied to the PIN diode problem (6.2.9). The results obtained will be compared to the results using the standard adaptive method.

The defect correction method is a method for solving semilinear finite element problems accurately and efficiently. The method involves solving one nonlinear problem and then a sequence of linear problems on finer grids. Although the theory in Chapter 4 is for *a priori* determined meshes, the meshes used in this section are obtained by *cautious* mesh refinement of a uniform mesh based on the *a posteriori* error estimate (6.2.11). This is an adaptive version of the defect correction method.

Before giving the algorithm for the method it is necessary to define the Fréchet derivative, $\mathrm{F}' : \mathcal{V}_h \to L\left(\mathcal{V}_h, (\mathcal{V}_h)'\right)$, of the nonlinear function, F, given by (6.2.9):

$$(\mathrm{F}'(\psi_h)v_h, w_h) := (\lambda^2 \nabla v_h, \nabla w_h) + \langle 2\delta^2 \cosh(\psi_h)v_h, w_h\rangle \quad v_h, w_h \in \mathcal{V}_h. \tag{6.2.12}$$

The adaptive defect correction algorithm is:

(1) Choose an initial finite element mesh $\mathcal{V}_h^0$ and a tolerance. Set k = 0.

(2) Seek a finite element solution, ψ_h^0, to (6.2.10) on the initial mesh. Solve the nonlinear problem using Newton's method.

(3) If the error in solution ψ_h^k, as measured by the *a posteriori* error estimate, is less than the tolerance then stop. Otherwise, refine the mesh *carefully* based on the *a posteriori* error estimate to obtain the new mesh.

(4) Solve the following linear problem on the current mesh:

$$\mathrm{F}'(\psi_h^k)e_h^{k+1} = -\mathrm{F}(\psi_h^k)$$

for e_h^{k+1}. Set $\psi_h^{k+1} = \psi_h^k + e_h^{k+1}$ and k = k+1. Return to step (3).

As before, the refinement is based on the L_2 *a posteriori* error estimate, but to conform with the defect correction theory in Chapter 4 the meshes are refined very cautiously (a maximum of 10% of the triangles are refined at each stage). The theory for the defect correction method also requires a sufficiently fine initial mesh. It was

found that, on average, the defect correction method required an initial mesh four times finer than the initial mesh for the standard method to converge.

It is known (e.g. [23]) that the finite element solution to (6.2.9) satisfies a discrete maximum principle, i.e. it only takes values between the boundary conditions (6.2.5) and (6.2.6):

$$\sinh^{-1}\left(\frac{-1}{2\delta^2}\right) \leq \psi_h(x) \leq \sinh^{-1}\left(\frac{1}{2\delta^2}\right) \quad x \in \Omega. \tag{6.2.13}$$

It was found that if too many of the triangles were refined at each refinement step of the algorithm then the discrete maximum principle (6.2.13) was violated (typically refining over 20% of the triangles would cause problems).

The defect correction method worked exceedingly well for the PIN diode problem (6.2.9) when λ was held fixed and $\delta \to 0$. This is probably because as δ decreases the layers in the solution become less severe.

δ^2	Initial mesh	Final number of mesh points	Final number of triangles	Number of refinement steps
1×10^{-5}	40×40	3745	7309	7
1×10^{-6}	40×40	3631	7083	7
1×10^{-7}	40×40	3734	7291	8
1×10^{-8}	40×40	3688	7199	7

Table 6.4: Results for the defect correction method for $\delta \to 0$ and $\lambda^2 = 1 \times 10^{-4}$. Results are for a tolerance of 5×10^{-3}.

The results for the defect correction method when $\delta \to 0$ and $\lambda^2 = 1 \times 10^{-4}$ are in Table 6.4. These results for the defect correction method compare very favourably with the results for the standard method in Table 6.1, particularly in terms of the number of iterations and mesh points the method requires for $\delta^2 = 1 \times 10^{-5}$. Figures 6-7 and 6-8 show the finite element solutions when $\delta^2 = 1 \times 10^{-5}$ and 1×10^{-8}, respectively. Comparing these with the finite element solutions produced by the standard method (Figures 6-2 and 6-3) shows that both methods produce comparable results. The only real difference between the solutions is that the defect correction solution has more mesh points in regions of slow change, this is due to the finer initial grid required for the method to work.

The PIN diode finite element problem, (6.2.9), when δ is held fixed and λ decreases is more difficult to solve using the defect correction method, mainly due to the size of

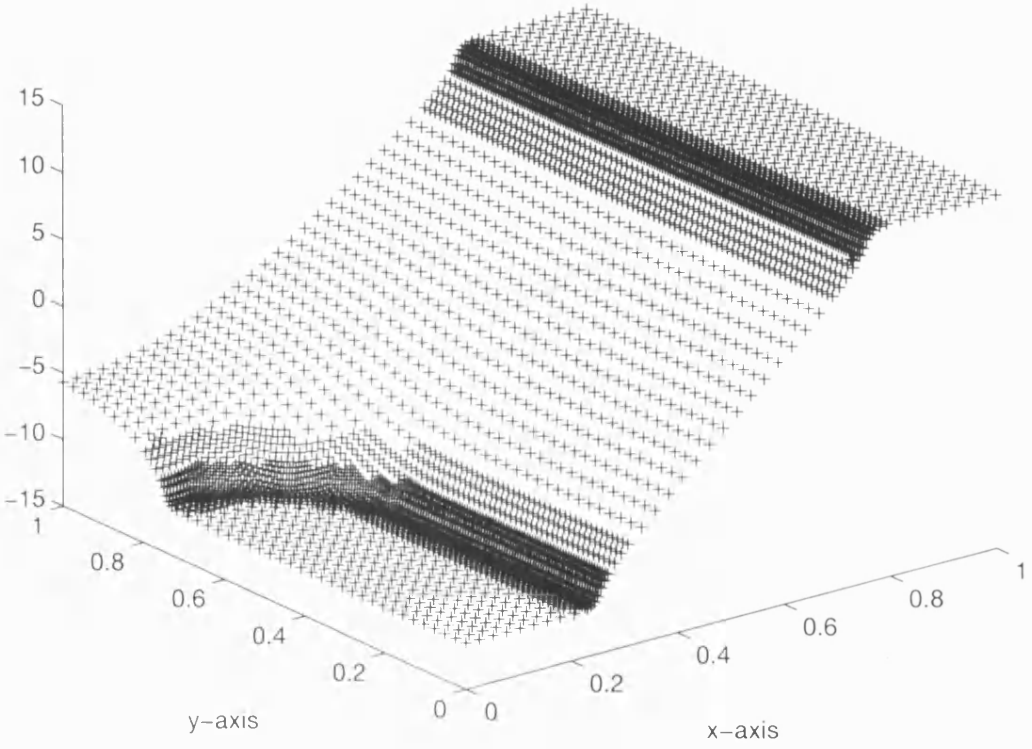


Figure 6-7: The defect correction finite element solution to the PIN diode problem when $\delta^2 = 1 \times 10^{-5}$ and $\lambda^2 = 1 \times 10^{-1}$.

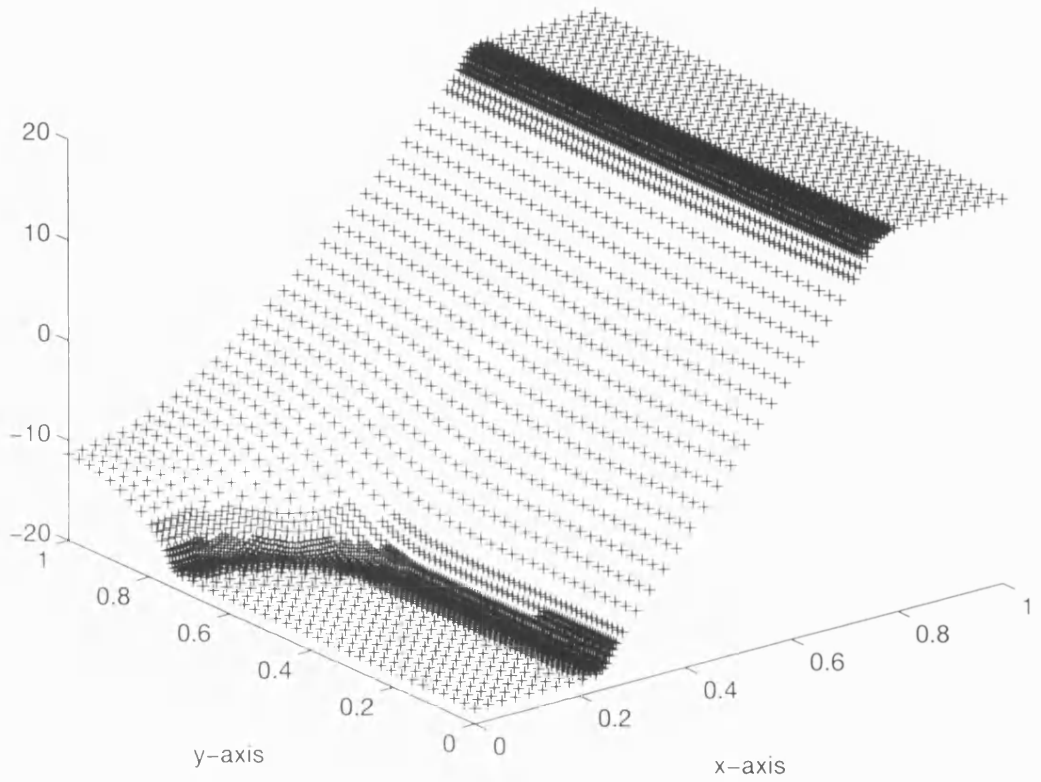


Figure 6-8: The defect correction finite element solution to the PIN diode problem when $\delta^2 = 1 \times 10^{-8}$ and $\lambda^2 = 1 \times 10^{-1}$.

the initial mesh required.

It was found that if a sufficiently fine initial mesh was not taken (6.2.13) was violated. For example the finite element solution to (6.2.9) with $\lambda^2 = 1 \times 10^{-6}$ and $\delta^2 = 1 \times 10^{-4}$ had an error of 65% for an initial mesh of size 40 × 40, the error decreased to 30% for an initial mesh of size 50 × 50 and the error was 12% for an initial mesh of size 60 × 60. Taking an initial mesh of size 70 × 70 for this problem is not desirable, partly because solving the initial nonlinear problem takes too long, but mainly because the mesh is over-refined in regions of slow change in the solution - the whole point of using the *a posteriori* error estimate is to avoid over-refinement.

Instead of using a very fine initial grid to start the defect correction method off when λ is small, a slightly different defect correction algorithm is used. The altered algorithm replaces step (4) of the original defect correction algorithm with the new step:

(4)′ Solve the following linear problem on the current mesh:

$$\mathrm{F}'(\psi_h^k)e_h^{k+1} = -\mathrm{F}(\psi_h^k)$$

for e_h^{k+1}. Set $\psi_h^{k+1} = \psi_h^k + e_h^{k+1}$. If the solution ψ_h^{k+1} exceeds the bounds (6.2.13) at a mesh point, then set the solution equal to the nearest boundary condition at that mesh point. Set k = k+1 and return to step (3).

Steps (1) to (3) of the new algorithm remains the same.

λ^2	Initial mesh	Final number of mesh points	Final number of triangles	Number of refinement steps
1×10^{-5}	40 × 40	3626	7077	9
1×10^{-6}	50 × 50	7022	13823	16
1×10^{-7}	60 × 60	10751	21239	22
1×10^{-8}	70 × 70	7617	14949	6

Table 6.5: Results for the altered defect correction method for $\lambda \to 0$ and $\delta^2 = 1 \times 10^{-4}$. Results are for a tolerance of 5×10^{-3}.

With this altered defect correction algorithm the method works very well on the PIN diode problem with δ fixed and λ decreasing. The results for the method are contained in Table 6.5 and in Figures 6-9 and 6-10. The solution produced by the defect correction method is identical to the solution produced by the standard adaptive method, except for the increased number of mesh points needed to obtain the same accuracy.

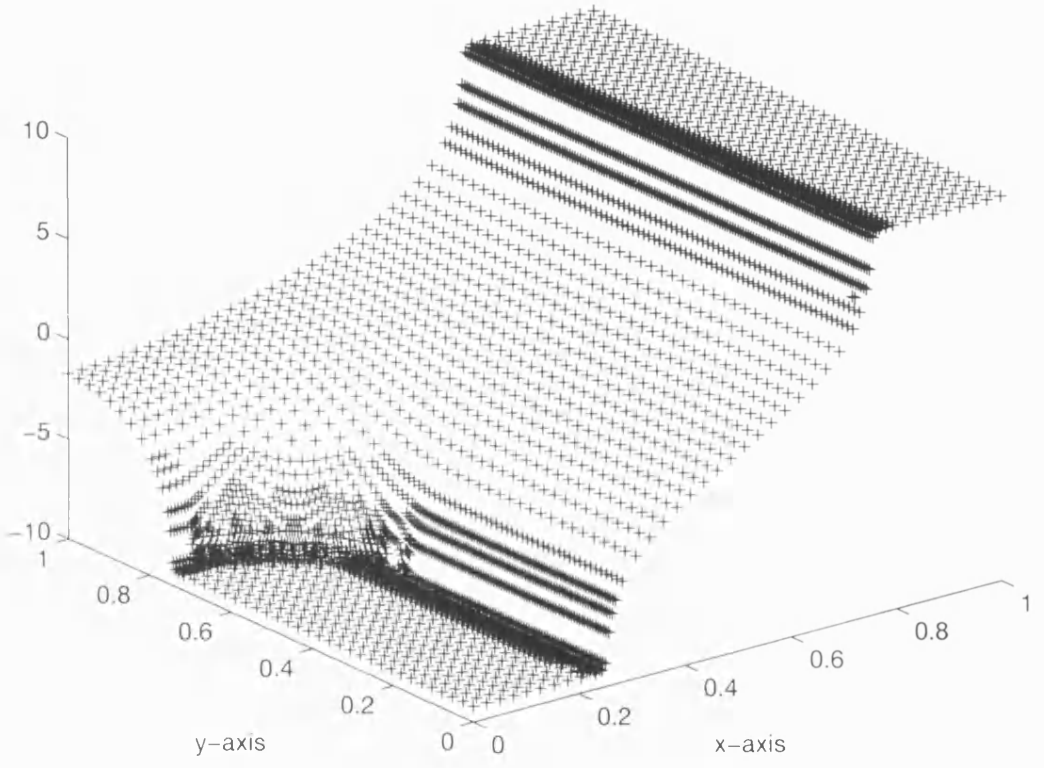


Figure 6-9: The defect correction finite element solution to the PIN diode problem when $\lambda^2 = 1 \times 10^{-5}$ and $\delta^2 = 1 \times 10^{-1}$. The solution is produced by using the altered defect correction algorithm.

It is interesting to compare the number of linear solves the methods require to find a finite element solution which has an error less than the given tolerance. A comparison of the total number of linear solves required for each method is given in Table 6.6. For the finite element PIN diode problems considered, the defect correction method requires fewer linear solves, despite the increased number of iterations for small λ. However, since the initial grids used in the defect correction method are much larger than those used in the standard adaptive method, the defect correction method still takes a comparable amount of time to solve the finite element problem.

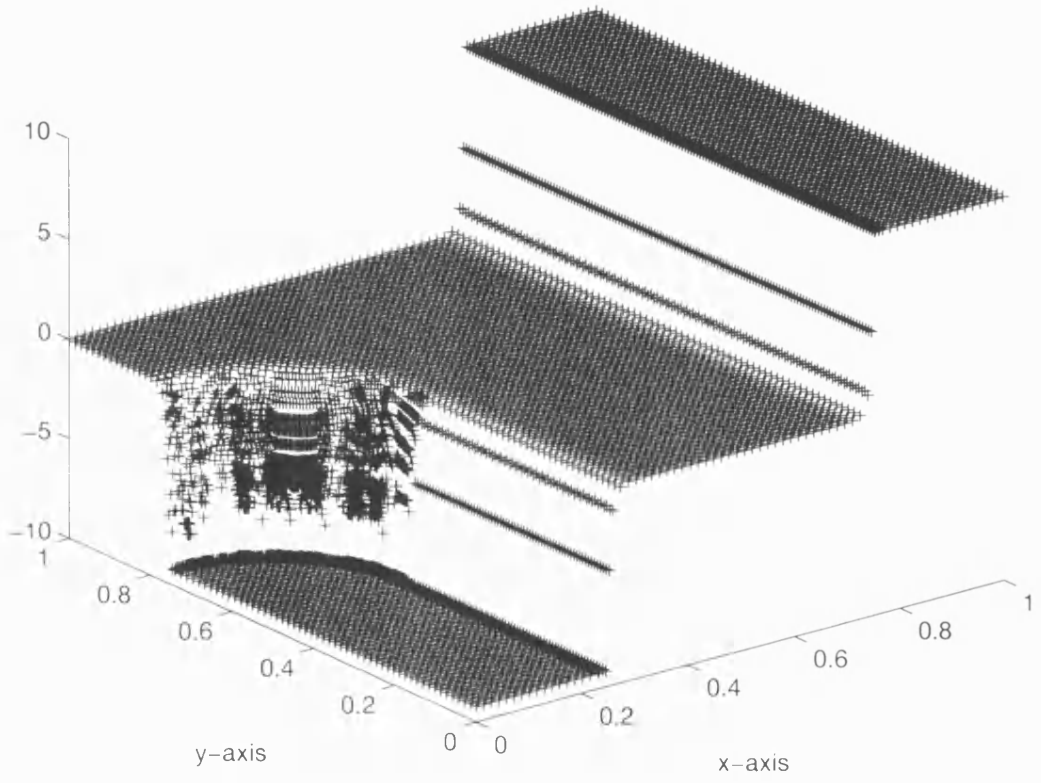


Figure 6-10: The defect correction finite element solution to the PIN diode problem when $\lambda^2 = 1 \times 10^{-8}$ and $\delta^2 = 1 \times 10^{-1}$. The solution is produced by using the altered defect correction algorithm.

In the time available it is not possible to optimise the code used to solve the finite element PIN diode problem with the standard and defect correction methods. It is therefore not possible to give cost profiles [the number of operations involved in solving a problem] for the methods. However, it is possible to get an idea of the relative costs of the methods. It is known that if a finite element problem was solved using an optimal

λ^2	δ^2	Number of linear solves for standard method	Number of linear solves for defect correction method
1×10^{-4}	1×10^{-5}	32	17
1×10^{-4}	1×10^{-7}	22	17
1×10^{-4}	1×10^{-8}	66	53
1×10^{-5}	1×10^{-4}	64	20
1×10^{-7}	1×10^{-4}	156	32
1×10^{-8}	1×10^{-4}	310	14

Table 6.6: The number of linear solves required to solve the PIN diode problem for different values of λ and δ using the standard adaptive and defect correction methods. The initial grids are those used before and change depending on the method, λ and δ. The tolerance is fixed at 5×10^{-3}.

method then the total cost would be:

$$\text{Number of mesh points} \times \text{Number of linear solves.}$$

Since the standard and defect correction methods considered here both involve adaptive procedures the estimates of the 'ideal' total costs are calculated using the formula:

$$\sum_{\text{Iterations}} \begin{matrix}\text{Number of mesh points} \\ \text{in the current grid}\end{matrix} \times \begin{matrix}\text{Number of linear} \\ \text{solves required}\end{matrix} .$$

The results in Table 6.7 suggest that the defect correction method is easily competitive in terms of cost. The only time the defect correction method has an 'ideal' cost significantly greater than the 'ideal' cost of the standard method is when $\lambda^2 = 1 \times 10^{-4}$ and $\delta^2 = 1 \times 10^{-8}$. The increased cost is due to the large initial mesh the defect correction method requires and the large number of linear solves it takes to solve the initial finite element problem to the required accuracy.

There appear to be two main problems with the defect correction method, both of these manifest themselves most clearly as $\lambda \to 0$. They are the large increase in the number of mesh points required and the increase in the number of iterations needed to obtain an accurate defect correction solution. The main reason for the increase in the number of mesh points required is the fine initial grid needed to start the method, but these grids are still not fine enough to capture the layers to the required tolerance. The second reason for the increase in the number of loops and mesh points is the cautious

λ^2	δ^2	Ideal cost for the standard method	Ideal cost for the defect correction method
1×10^{-4}	1×10^{-5}	39136	35497
1×10^{-4}	1×10^{-7}	33592	35789
1×10^{-4}	1×10^{-8}	61304	83445
1×10^{-5}	1×10^{-4}	81621	42419
1×10^{-7}	1×10^{-4}	357578	198885
1×10^{-8}	1×10^{-4}	655109	74619

Table 6.7: The ideal cost of finding an accurate finite element solution to the PIN diode problem for different values of λ and δ using the standard adaptive and defect correction methods. The tolerance is fixed at 5×10^{-3}.

adaptive strategy, new mesh points are mainly introduced within the layers at each iteration of the algorithm, but extra mesh points are also needed at the outer parts of each of the layers, resulting in a larger than expected *a posteriori* error estimate. It is only when the errors at the extreme points of a layer starts to compete with the error from within the layer that the *a posteriori* error estimate starts to reduce and the algorithm terminate.

Despite the extra loops and mesh points required the defect correction method still competes well with the standard adaptive method in terms of 'ideal' cost and accuracy achieved. Even with a large increase in the number of iterations required the defect correction method requires a smaller number of linear systems to be solved. The results contained in this section suggest that the defect correction method is an exceedingly good method, providing care is taken with the refinement procedure and the initial grid is sufficiently fine.

6.3 Simplified MOSFET

The MOSFET (Metal Oxide Semiconductor Field Effect Transistor) is one of the most important semiconductor devices since it can be used as a switch without consuming any power. There have been numerous numerical models of the MOSFET device, for example [62], [27], [34] and [52]. Here we solve a reduced model proposed in [51] in which some simplifying assumptions enable us to model the device and capture its most important features, yet allow us to focus on the most important part of the device and reduce the computational effort needed for the simulation. This simplified model is a

2×2 coupled system, with variables $\tilde{\psi}$ and $\tilde{v}$, comprising scaled electrostatic potential and electron quasi-Fermi level respectively ((6.3.18)-(6.3.19) below). The efficient application of adaptivity to this system is a challenging open problem: One has to choose between using one the variables as a basis for determining the adaptive meshes - usually $\tilde{\psi}$, see [47] - or trying to adapt on both variables which requires more computations as in [27]. In this thesis we have only used the former approach. As one shall see below, although this is sometimes successful there may be some situations in which it is unsatisfactory. It is still an open question to investigate fully adaption of meshes for different components of the solution of a coupled system which may have difficulties in different places.

The MOSFET has four contacts: the source, drain, bulk and gate contacts. The source and drain are connected to highly doped n-type regions of semiconductor and the bulk is connected to a p-type region with much lower doping. Between the source and drain is a thin layer of oxide (silicon dioxide for a silicon based semiconductor) and a metal contact, the gate. A typical MOSFET arrangement is shown in Figure 6-11.

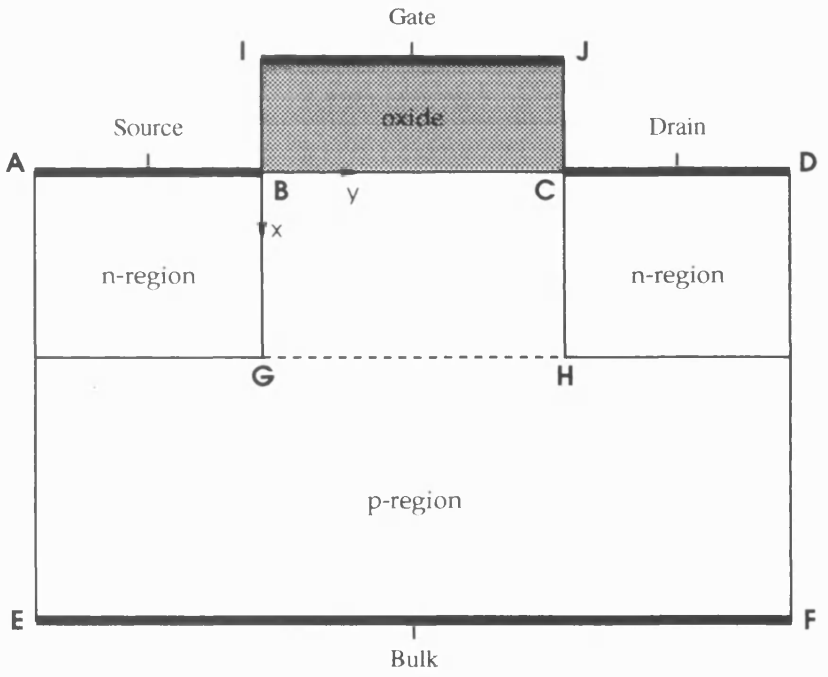


Figure 6-11: Cross section of a simplified MOSFET device

When a sufficiently large voltage is applied at the gate an inversion layer forms near the semiconductor/oxide interface (BC in Figure 6-11). In this area the electron density

dominates the hole density and so this region near the semiconductor/oxide interface is known as an *n-channel.* When there is a sufficiently large voltage difference between the two contacts the n-channel is able to carry a significant current from source to drain. This current can be switched on and off by applying different voltages at the gate. The aim of this section will be to numerically model how the electrons behave in this channel for different applied voltages.

The n-type region below the source contact will be referred to as the source region. Similarly the n-type region below the drain contact will be called the drain region. The remaining p-type region will be referred to as the bulk region.

The drift-diffusion equations (in the quasi-Fermi variables) introduced in Section 1.3 hold in the semiconductor region, Ω_s (ADFE in Figure 6-11). To simplify the model we assume that the generation/recombination rate is set to zero. The oxide region, Ω_{ox} (IJCB in Figure 6-11), is assumed to be free of charge ($n = p = 0$) and here Laplace's equation holds for the electrostatic potential ψ. The equations modelling the MOSFET can therefore be written as:

$$-\lambda^2\Delta\psi + \delta^2\{\exp(\psi - v) - \exp(w - \psi)\} = d, \text{ in } \Omega_s, \tag{6.3.14}$$

$$-\nabla.(\exp(\psi - v)\nabla v) = 0, \text{ in } \Omega_s, \tag{6.3.15}$$

$$\nabla.(\exp(w - \psi)\nabla w) = 0, \text{ in } \Omega_s, \tag{6.3.16}$$

$$\Delta\psi = 0, \text{ in } \Omega_{ox}. \tag{6.3.17}$$

As the voltage applied at the source contact acts as a reference voltage for the drain voltage we set the source voltage to be zero. It is not necessary to always apply a voltage to the bulk contact and for the purposes of our simulation the applied voltage at the bulk will also be assumed to be zero. Defining the voltage applied at the drain contact to be V_d and the voltage applied at the gate to be V_g, the Dirichlet boundary conditions on the system can be written as:

$$v = w = 0, \quad \psi = \sinh^{-1}(+1/2\delta), \quad \text{at the source contact,}$$

$$v = w = 0, \quad \psi = \sinh^{-1}(-1/2\delta), \quad \text{at the bulk contact,}$$

$$v = w = V_d/U_T, \quad \psi = \sinh^{-1}(+1/2\delta) + V_d/U_T, \quad \text{at the drain contact,}$$

$$\psi = \sinh^{-1}(-1/2\delta) + V_g/U_T, \quad \text{at the gate contact,}$$

where U_T is the thermal voltage given in Chapter 1. At all other boundaries of the device homogeneous Neumann conditions hold for ψ, v and w (where appropriate). All that remains to specify is the interface condition for ψ at the join between the oxide and semiconductor. Taking the origin of the coordinate system for the model at the point B in Figure 6-11 and with the x- and y-axes as shown, we impose the usual interface conditions for ψ:

$$\psi(0-, y) = \psi(0+, y), \quad \epsilon_{ox}\partial_x\psi(0-, y) = \epsilon_s\partial_x\psi(0+, y),$$

where ϵ_{ox} is the absolute permittivity of the oxide and ϵ_s is the absolute permittivity of the silicon semiconductor. Values for these quantities are given in Chapter 1. This condition forces the potential and vertical component of the electric displacement ($\epsilon\nabla\psi$) to be continuous across the interface.

The properties of the MOSFET with various applied voltages are studied in Chapter 3 of [65]. With $V_g = 0$ it is known that there are very few electrons in the channel. Non-zero gate voltage V_g creates an electric field near the semiconductor/oxide interface, this repels holes and induces electrons into the channel. As the gate voltage increases more electrons flow into the channel, increasing the current. However increasing the drain voltage V_d has the affect of repelling electrons from the drain end of the channel and reduces the density of electrons in this region.

Since we are interested in modelling the behaviour of the MOSFET in the n-channel (near the semiconductor/oxide interface) we introduce a simplified model which focuses on the region BCHG in Figure 6-11. This model is also discussed in [51, Section 4.7].

6.3.1 The Simplified MOSFET Model

This section describes the reduction of the MOSFET model to a boundary value problem in the small region represented by BCHG in Figure 6-11.

In [51, Section 4.7] (using singular perturbation analysis as $\lambda \to 0$) it is shown that v is approximately equal to the scaled applied source and drain voltages in the source and drain regions, respectively. In the bulk region w is approximately equal to the applied voltage at the bulk contact. With the given applied voltages we therefore assume:

$$v = 0 \text{ in the source region,}$$

$$v = V_d/U_T \text{ in the drain region,}$$
$$w = 0 \text{ in the bulk region.}$$

Defining $\tilde{l}$ to be the oxide thickness (the length of the segment IB in Figure 6-11) and assume that the channel length (the length of the segment BC in Figure 6-11) is large in comparison. This is reasonable since a typical channel lengths is 250 nm, while a typical oxide thickness is 4.5 nm. Introduce the independent variable $\xi = x/\tilde{l}$. In this new variable the oxide region is transformed to a unit square and the potential satisfies

$$\partial_\xi^2 \psi + \tilde{l}^2 \partial_y^2 \psi = 0.$$

In the limit as $\tilde{l} \to 0$, this potential equation becomes a one dimensional problem which, bearing in mind the conditions at the semiconductor/oxide interface, leads to the interface boundary condition ([51, Section 4.7]):

$$\frac{\epsilon_s \tilde{l}}{\epsilon_{ox}} \partial_x \psi = \psi - \sinh^{-1}\left(\frac{-1}{2\delta^2}\right) - \frac{V_g}{U_T}, \text{ on the interface BC.}$$

Since the hole concentration, w, is known to be approximately equal to zero in the region BCHG in Figure 6-11 our simplified model for this region does not need to include the hole continuity equation.

Define $\gamma = (\log(\tilde{d}/n_i))^{-1}$ and $\tilde{\lambda} = \lambda/(\gamma)^{\frac{1}{2}}$, where $\tilde{d}$ is the maximum of the doping profile and n_i is the intrinsic concentration. Then introducing the rescaled variables:

$$\tilde{\psi} = \gamma\psi, \quad \tilde{v} = \gamma v, \quad \xi = x/\tilde{\lambda},$$

we arrive at the simplified MOSFET system:

$$\partial_\xi^2 \tilde{\psi} + \tilde{\lambda}^2 \partial_y^2 \tilde{\psi} = \exp\left(\frac{\tilde{\psi} - \tilde{v} - 1}{\gamma}\right) - \exp\left(\frac{-\tilde{\psi} - 1}{\gamma}\right) + 1, \tag{6.3.18}$$

$$\partial_\xi\left(\exp\left(\frac{\tilde{\psi} - \tilde{v} - 1}{\gamma}\right)\partial_\xi \tilde{v}\right) + \tilde{\lambda}^2 \partial_y\left(\exp\left(\frac{\tilde{\psi} - \tilde{v} - 1}{\gamma}\right)\partial_y \tilde{v}\right) = 0. \tag{6.3.19}$$

Defining $\tilde{V}_g = \gamma V_g/U_T$ and $\tilde{V}_d = \gamma V_d/U_T$ to be the rescaled gate and drain voltages, [51] derives the following boundary conditions on the new variables $\tilde{\psi}$ and $\tilde{v}$ in the region

BCHG of Figure 6-11:

$$\begin{aligned}
\alpha\partial_\xi\tilde{\psi} &= \tilde{\psi} - \gamma\sinh^{-1}(-1/2\delta^2) - \tilde{V}_g, && \text{on BC,}\\
\tilde{\psi} &= -1 - \gamma\log\frac{1}{2}\left(1+\left(1+4\exp\left(\frac{-\tilde{v}-2}{\gamma}\right)\right)^{\frac{1}{2}}\right), && \text{on GH,}\\
\tilde{v} &= 0, && \text{on BG,}\\
\tilde{v} &= \tilde{V}_d, && \text{on CH,}\\
\partial_\xi\tilde{v} &= 0, && \text{on BC and GH.}
\end{aligned}$$

In the above $\alpha = \epsilon_s\tilde{l}/\epsilon_{ox}\tilde{\lambda}$. Since [51] does not specify the boundary conditions for $\tilde{\psi}$ on BG and CH of the reduced domain (these boundary conditions are shown not to affect the result) we take homogeneous Neumann boundary conditions for ψ on these boundaries.

Remark 6.3.1 *The boundary conditions for $\tilde{v}$ are natural given the assumptions we have made. They suggest that current flow in the device is only in the direction tangential to the semiconductor/oxide interface, this fits in with our idea that the current flows along the n channel from source to drain. However, numerical simulations in [62] suggest that this is not exactly the case for real devices.*

6.3.2 Numerical Simulations for the Simplified MOSFET Model

In this section finite element solutions to the simplified MOSFET system (6.3.18) and (6.3.19), with the given boundary conditions, are found for a variety of applied voltages. We use an adaptive method for (6.3.18)-(6.3.19) where the grids are refined using the L_2 *a posteriori* error estimate for $\tilde{\psi}$ obtained in Chapter 5. This problem is considerably more challenging than the semilinear problems encountered in Section 6.2, mainly due to the exponential coefficient in (6.3.19). For this reason we have not attempted to apply the adaptive version of the defect correction method of Chapter 4 but instead we simply solve the full nonlinear system corresponding to each refinement step to full accuracy using a variant of Gummel's method (Section 3.3.3).

The channel length is taken to be 250 nm and the oxide thickness is taken as 4.5 nm. To simplify our calculations we took $\tilde{\lambda} = 1$. This value of $\tilde{\lambda}$ corresponds to a germanium semiconductor with a maximum doping profile of approximately 6×10^{15} (germanium has a relative permittivity of 16.1 as opposed to 11.7 for silicon). Different devices will

yield values of $\tilde{\lambda}$ which are not equal to one, the methods described here should work well even in those cases unless $\tilde{\lambda}$ becomes very close to zero.

Our adaptive procedure is based purely on the *a posteriori* error estimate for $\tilde{\psi}$ obtained from the semilinear equation (6.3.18), with $\tilde{v}$ considered to be known (we use the current value of $\tilde{v}$). In [27] it is shown that for devices in which the current flows predominantly perpendicular to the layer (e.g. a PN diode) an error estimate which is based only on the Poisson equation is sufficient for accurate refinement. However, for devices in which the current is parallel to the layer (e.g. a MOSFET diode) an error estimate based on all three semiconductor equations is in general desirable. In this special case we are only interested in the electron concentration near the n-channel of the MOSFET and so we believe it is sufficient to base our refinement strategy on the Poisson equation [since the electrostatic potential $\tilde{\psi}$ varies most rapidly in this channel].

Our refinement strategy uses the L_2 *a posteriori* error estimate with estimated constant, as discussed in Chapter 5. This error estimate depends on α such that $\tilde{\psi} \in H^{1+\alpha}$. In Chapter 4 we calculated α in the case of a polygonal domain and mixed Dirichlet and Neumann boundary conditions. However in this case we have the domain BCHG with Neumann conditions on BG and CH, Dirichlet conditions on GH and Robin conditions on BC. Thus we cannot automatically use the estimate of α from Chapter 4. However we observe that essentially α is computed by finding the regularity of solutions to Laplace's equation subject to these boundary conditions. Therefore we need to examine the regularity near the collision points between Neumann and Robin conditions in BCHG (i.e. at B and C). Without loss of generality we restrict to B and follow (formally) the procedure in Grisvard ([35, pages 49-51]). We take polar coordinates about B and seek a solution of Laplace's equation in the form

$$\psi = r^{\alpha}(\cos(\alpha\theta) + i\sin(\alpha\theta)). \tag{6.3.20}$$

It turns out that α should satisfy the nonlinear equation $\alpha = -1/\tan(\alpha\pi/2)$ which has solution $\alpha = 1.654$. So $\tilde{\psi} \in H^{1+\alpha}$ near B and C with α approximately equal to 1.654, but we also know $\tilde{\psi} \in H^2$ near G and H from Chapter 4, so in our L_2 *a posteriori* error estimate we put $\alpha = 1$.

Our initial guess for $\tilde{v}$ is the plane joining the two Dirichlet boundary conditions for $\tilde{v}$. For $\tilde{\psi}$ we interpolate the plane connecting the Dirichlet boundary condition at

$\xi = -1$ and the condition $\gamma \sinh^{-1}(-1/2\delta^2) - \tilde{V}_g$ at the boundary $\xi = 0$ (this condition arises from the Robin boundary condition at the semiconductor/oxide interface).

The nonlinear systems are solved using the following adaptive variant of Gummel's method [for clarity we write the algorithm in the 'natural', rather than the finite element, variables, the actual scheme used operates on the finite element approximations]:

(1) Set up an initial coarse mesh and initial guesses to $\tilde{\psi}$ and $\tilde{v}$: $\tilde{\psi}_0^0, \tilde{v}_0^0$. Set $l = 0$.

(2) For $k = 0, 1, \ldots$

(a) Solve for $\tilde{\psi}_{k+1}^l$ using Newton's method:

$$\Delta\tilde{\psi}_{k+1}^l = \exp\left(\frac{\tilde{\psi}_{k+1}^l - \tilde{v}_k^l - 1}{\gamma}\right) - \exp\left(\frac{-\tilde{\psi}_{k+1}^l - 1}{\gamma}\right) + 1.$$

(b) Solve for $\tilde{v}_{k+1}^l$:

$$\nabla.\left(\exp\left(\frac{\tilde{\psi}_{k+1}^l - \tilde{v}_k^l - 1}{\gamma}\right)\nabla\tilde{v}_{k+1}^l\right) = 0.$$

(c) Repeat until

$$\max\left\{\|\tilde{\psi}_{k+1}^l - \tilde{\psi}_k^l\|_\infty, \|\tilde{v}_{k+1}^l - \tilde{v}_k^l\|_\infty\right\} < 1 \times 10^{-6}.$$

The linear systems are solved using GMRES with ILU decomposition as a preconditioner.

(3) Define the computed values of $\tilde{\psi}$ and $\tilde{v}$ obtained in step (2) to be $\tilde{\psi}_0^{l+1}, \tilde{v}_0^{l+1}$, the initial guesses for the next outer iteration. Set $l = l + 1$.

(4) If the L_2 *a posteriori* error estimate for $\tilde{\psi}$ is less than 5×10^{-3} then stop. Otherwise refine the mesh based on the *a posteriori* estimate as discussed in Chapter 5 and return to (2).

Remark 6.3.2 *The linear system in step (2b) of the algorithm is discretised by averaging the coefficient* $\exp((\tilde{\psi}_{k+1}^l - \tilde{v}_k^l - 1)/\gamma)$ *on each triangle and treating the operator as* $\nabla.(a\nabla\cdot)$, *where* a *is constant on each triangle. The resulting discretised equation is diagonally scaled as the exponential term varies significantly in size and the resulting discretisation often has blocks of entries which are too small to allow accurate solution.*

This adaptive scheme should produce accurate finite element approximations to the scaled variable $\tilde{\psi}$ and less accurate approximations to $\tilde{v}$. However we are mainly interested in the electron concentration n, this is obtained from $\tilde{\psi}$ and $\tilde{v}$ by the formula (see Section 1.3):

$$n = n_i \exp\left(\frac{\tilde{\psi} - \tilde{v}}{\gamma}\right),$$

where n_i is the intrinsic concentration (given in Table 1.1).

Various electron concentrations are shown in Figures 6-12, 6-13, 6-14, 6-15 and 6-16 for a variety of gate and drain voltages. A typical picture of the finite element approximation to the scaled electrostatic potential $\tilde{\psi}$ is shown in Figure 6-17. Figure 6-18 shows a typical picture of the scaled electron quasi-Fermi level $\tilde{v}$.

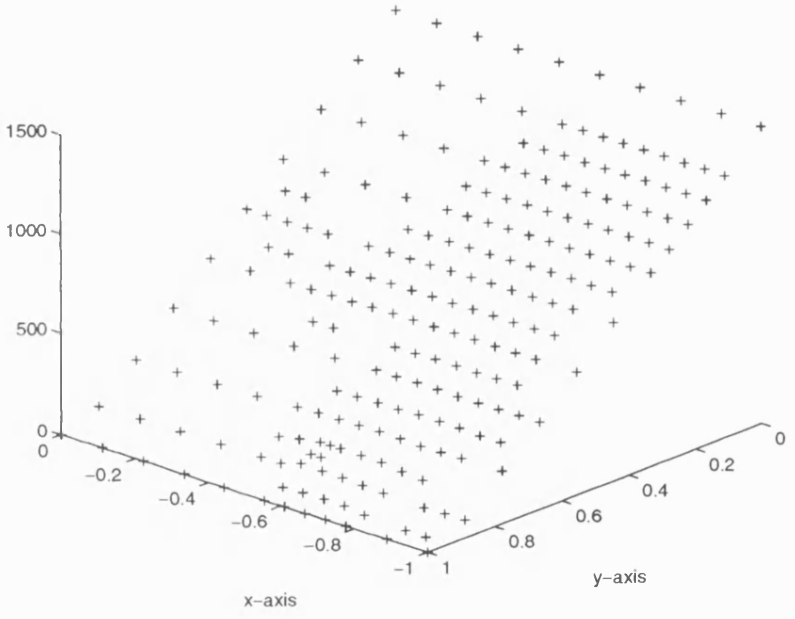


Figure 6-12: The finite element approximation to the electron concentration in the simplified MOSFET region when $V_g = 0.0$ and $V_d = 0.2$.

Figure 6-12 shows that when there is no voltage applied at the gate contact there are very few electrons in the channel region. Note: a typical number of electrons would be of order 10^{19}, rather than the order of 10^3 seen in Figure 6-12 when there is no gate voltage. Comparing Figure 6-12 with Figures 6-13 and 6-15 shows that increasing the gate voltage dramatically increases the number of holes and electrons in the channel. This conforms with the predictions in [65].

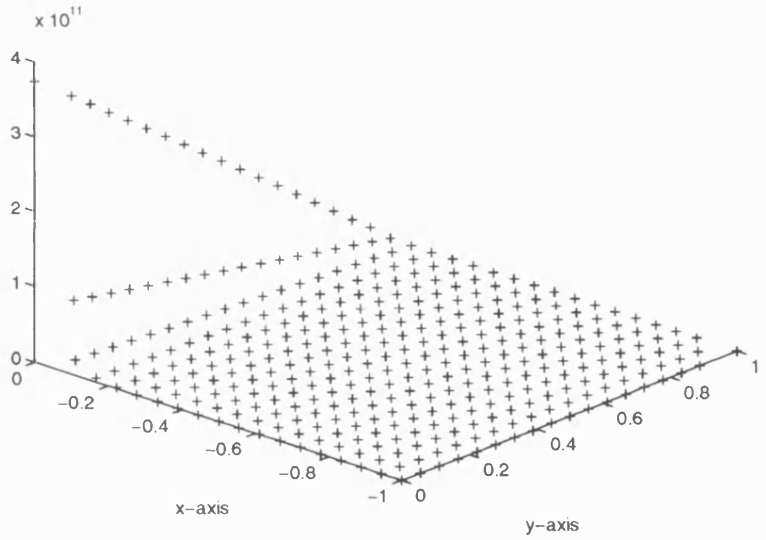


Figure 6-13: The finite element approximation to the electron concentration in the simplified MOSFET region when $V_g = 0.5$ and $V_d = 0.1$.

Comparing Figure 6-13 with Figure 6-14 and Figure 6-15 with Figure 6-16 we notice that, as the voltage applied at the drain contact increases, the concentration of electrons at the drain end of the model ($y = 1$ in the domain) decreases, which is consistent with the physics of the device discussed in [65].

In all these experiments $\tilde{\psi}$ has a gentle slope between the value of the Dirichlet boundary condition and $\gamma \sinh^{-1}(-1/2\delta^2) - \tilde{V}_g$ (at the semiconductor/oxide interface). Adaption is not really needed for $\tilde{\psi}$ for most gate and drain voltages considered, however we force the program to do a minimum of two outer iterations so we can estimate the constant appearing in the *a posteriori* error estimate. For the applied voltages considered in Figure 6-15 we forced the code to do five outer iterations. The new mesh points were introduced in the channel region near the semiconductor/oxide interface, supporting the view that, for the simplified MOSFET, refinement based only on the Poisson equation is sufficient to capture all the important detail in the channel.

Figure 6-18 shows the scaled electron quasi-Fermi level when the gate voltage is 0.5 volts and the drain voltage is 0.2 volts. The picture shows that $\tilde{v}$ has a sharp layer at the interface between the bulk and drain regions and a slight upward slope near the interface between the semiconductor and the oxide. Our adaptive procedure does

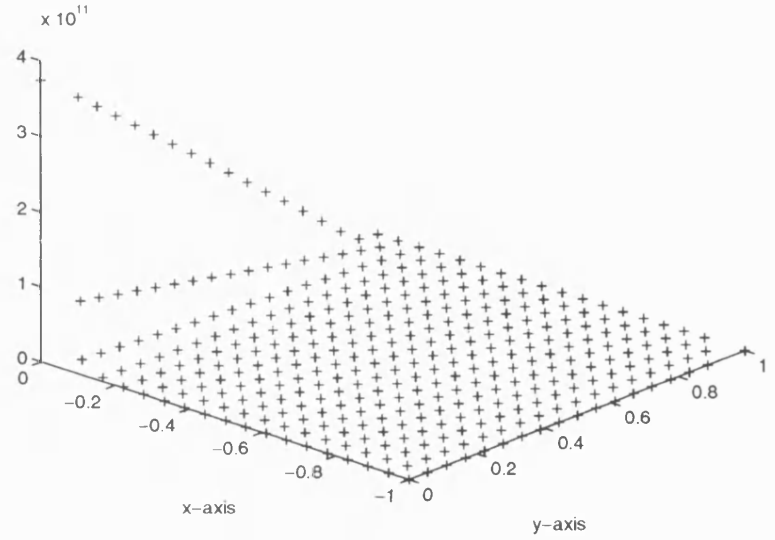


Figure 6-14: The finite element approximation to the electron concentration in the simplified MOSFET region when $V_g = 0.5$ and $V_d = 0.5$.

not capture the layer behaviour in $\tilde{v}$ accurately, to do this we would need an adaptive procedure based on *a posteriori* error estimates for both ψ and v (as suggested in [27]). An adaptive procedure for both ψ and v is outside the scope of this thesis, however some work on this problem has been considered in [27] and [18].

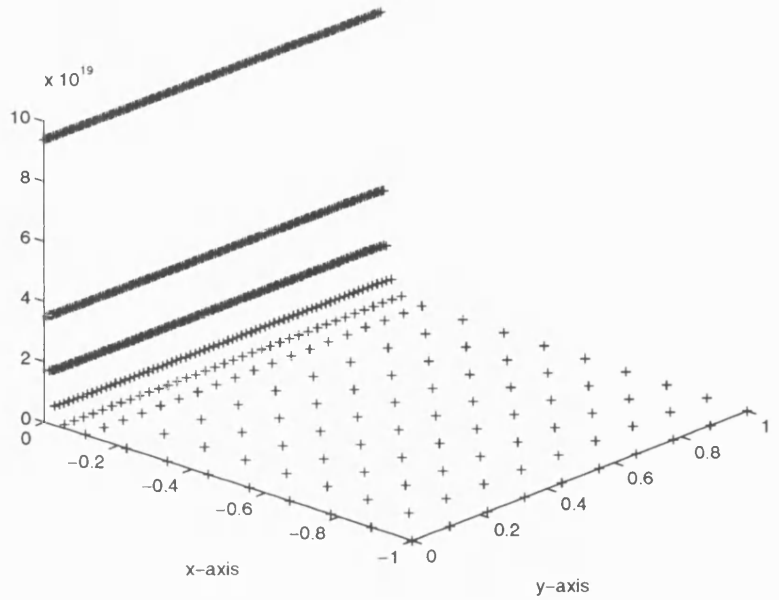


Figure 6-15: The finite element approximation to the electron concentration in the simplified MOSFET region when $V_g = 1.0$ and $V_d = 0.0$.

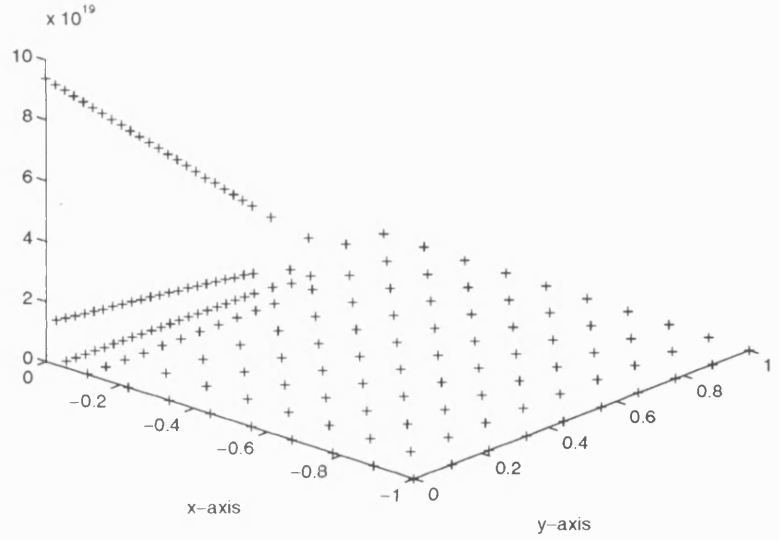


Figure 6-16: The finite element approximation to the electron concentration in the simplified MOSFET region when $V_g = 1.0$ and $V_d = 0.5$.

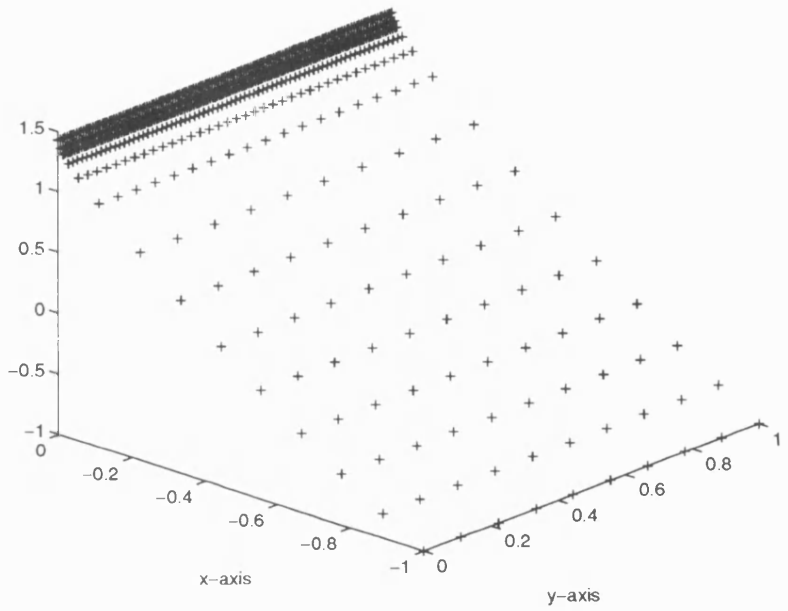


Figure 6-17: The finite element approximation to the scaled electrostatic potential, $\tilde{\psi}$, in the simplified MOSFET region when $V_g = 1.0$ and $V_d = 0.0$.

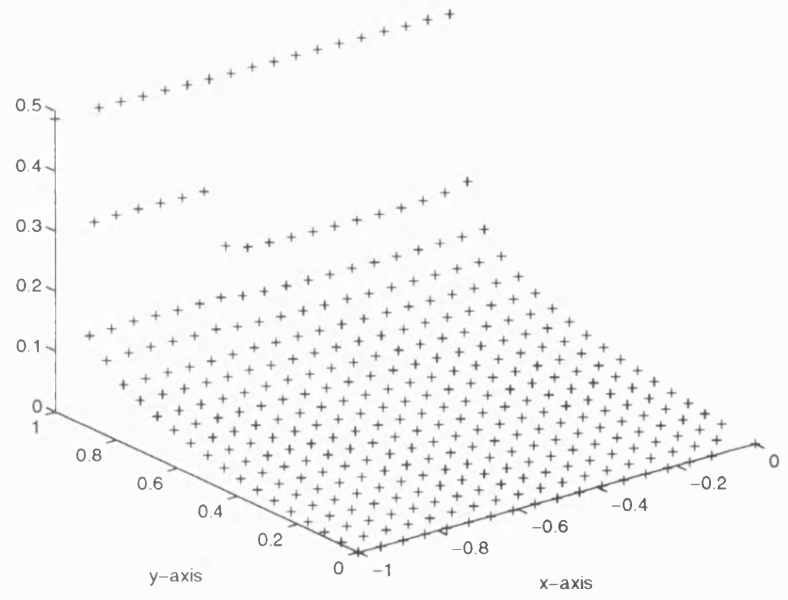


Figure 6-18: The finite element approximation to the scaled quasi-Fermi variable $\tilde{v}$ in the simplified MOSFET region when $V_g = 0.5$ and $V_d = 0.2$.

Appendix A

Matrix Theory

In this appendix we define some terms that will be used repeatedly in other chapters.

A.1 Graph Theory

Graph theory is a very powerful tool and is often used to find out the properties of a matrix. For example it is possible to use graph theory to deduce if a matrix is non-singular. Information can be found in a variety of books, particularly Varga ([69]) and Hackbusch ([37]). In this section a general overview will be provided and most of the terms used will be taken from [69].

Definition A.1.1 *Let $A = (a_{i,j})$ be an $n \times n$ matrix and let I denote the set $1, 2, \ldots, n$. The graph, $\mathcal{G}(A)$, of A is a subset of all pairs from $I \times I$ and is given by:*

$$\mathcal{G}(A) = \{(i,j) \in I \times I : a_{i,j} \neq 0\}$$

An index $i \in I$ is called a node. Node i is said to be *directly connected* to node j if the entry $a_{i,j}$ of A is non-zero. The set $\mathcal{G}(A)$ can thought of in the following visual way: If $a_{i,j}$ is non-zero represent the connection from i to j by means of an arrow pointing from i to j (this arrow is a path from i to j: $P_i P_j$). If $a_{i,j}$ and $a_{j,i}$ are both nonzero then the node i is directly connected to node j and node j is directly connected to node i. An example of the graph of a matrix can be found in Figure A-1.

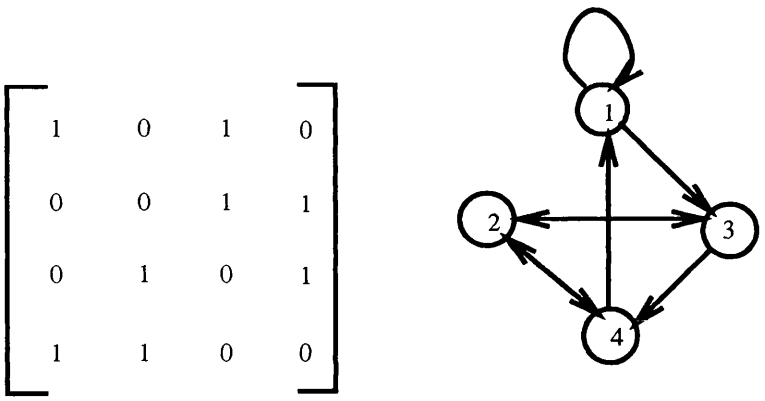


Figure A-1: An example of a matrix and its graph. Node 1 is directly connected to node 3 and is connected, but not directly connected, to node 2.

Node i is said to be *connected* to node j if there exists a series of direct connections linking i and j, i.e. there exists a series of paths connecting node i to node j:

$$P_i P_{i_1},\ P_{i_1} P_{i_2},\ \ldots, P_{i_n} P_j,$$

where $i_1, i_2, \ldots, i_n$ are nodes in I.

Remark A.1.2 *If A is symmetric then i is (directly) connected to j if and only if j is (directly) connected to i.*

Definition A.1.3 *A matrix is called connected if, for any two nodes i and j, i is connected to j. For each node i, we denote the set of nodes which i is connected to by $\mathcal{G}_i$, i.e. $\mathcal{G}_i := \{j : i \text{ is connected to } j\}$.*

A.2 Miscellaneous Results and Definitions

The following definitions and theorems are from Varga [69]. They will be used in conjunction with the definitions in the previous section to show various properties of the matrices arising from the discretisation of the semiconductor system.

Definition A.2.1 *An $n \times n$ matrix $A = (a_{i,j})$ is called diagonally dominant if:*

$$|a_{i,i}| \geq \sum_{j=1, j\neq i}^{n} |a_{i,j}| \tag{A.2.1}$$

for all $i \in \{1, 2, \ldots, n\}$. *A is called strictly diagonally dominant if the strict inequality in (A.2.1) is valid for all* $i \in \{1, 2, \ldots, n\}$.

Definition A.2.2 *An* $n \times n$ $(n \geq 2)$ *matrix A is reducible if there exists an* $n \times n$ *permutation matrix P such that*

$$PAP^T = \begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{bmatrix} \tag{A.2.2}$$

where $A_{1,1}$ *is an* $r \times r$ *submatrix and* $A_{2,2}$ *is an* $(n-r) \times (n-r)$ *submatrix* $(1 \leq r < n)$. *If no such permutation matrix exists then A is called irreducible. A* 1×1 *matrix is irreducible if its single entry is non-zero and reducible otherwise.*

Theorem A.2.3 (Theorem 1.6 of [69])
An $n \times n$ *matrix is irreducible if its graph is connected.*

Example The matrix

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$$

is reducible, but the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

is irreducible since it is connected.

Definition A.2.4 *An* $n \times n$ *matrix* $A = (a_{i,j})$ *is irreducibly diagonally dominant if it is irreducible, diagonally dominant and has at least one row, row i say, such that*

$$|a_{i,i}| > \sum_{j=1, j \neq i}^{n} |a_{i,j}|.$$

An irreducibly diagonally dominant matrix has some important properties, for instance such a matrix will be non-singular ([37]) and if it satisfies certain sign properties the matrix will have a positive inverse ([69]). However the condition that the matrix should be irreducible is quite difficult to satisfy in practice. Hackbusch (in [37]) weakens this requirement by making the following definition:

Definition A.2.5 *An $n \times n$ matrix $A = (a_{i,j})$ is essentially diagonally dominant if A is diagonally dominant and if for each $i \in \{1, 2, \ldots, n\}$, there exists a $k \in \{1, 2, \ldots, n\}$, such that i is connected to k (i.e. $k \in \mathcal{G}_i$) and*

$$|a_{k,k}| > \sum_{j=1, j \neq k}^{n} |a_{k,j}|.$$

Remark A.2.6 *For irreducible matrices, essentially and irreducibly diagonally dominant are equivalent.*

It is shown in [37] that an essentially diagonally dominant matrix is non-singular. It is also known ([37, Theorem 6.4.10]) that an essentially diagonally dominant matrix satisfying certain sign conditions will have a non-negative inverse (rather than a positive inverse as is the case with irreducibly diagonally dominant matrices).

The next theorem shows that if two matrices are sufficiently close in norm and one is non-singular then both are non-singular.

Theorem A.2.7 (Theorem 3.1.4 of [28])
Let A and B be any two square (real) matrices and $\| \cdot \|$ be any (real) matrix norm. If A is nonsingular and $\| A^{-1}(B - A) \| < 1$, then B is nonsingular and

$$\| B^{-1} \| \leq \frac{\| A^{-1} \|}{1 - \| A^{-1}(B - A) \|}. \tag{A.2.3}$$

Appendix B

Mass lumping

In the finite element method in this thesis we often replace Galerkin approximations to zero-order terms by their mass lumped versions. The purpose of this appendix is to introduce mass lumping in the context of a certain discrete bilinear form.

As in previous chapters consider a domain $\Omega \subset \mathbb{R}^2$. Define a triangulation of the domain $\mathcal{T}_h = \{T\}$ such that $\Omega = \bigcup_{T \in \mathcal{T}_h} T$. We call each of the vertices of the triangles a mesh point. Let h denote the maximum diameter of the triangles in the triangulation and $\mathcal{N}$ denote the set of mesh points of Ω. Define $\mathcal{V}_h$ to be the space of piecewise linear functions v such that v is continuous on Ω and $v|_T$ is linear for each $T \in \mathcal{T}_h$.

A basis for the piecewise linear finite element space $\mathcal{V}_h$ can be described in the following way: For each mesh point $p \in \mathcal{N}$ define ϕ_p to be a function such that $\phi_p(q) = \delta_{pq}$, $q \in \mathcal{N}$, where δ_{pq} is the Kronecker delta. ϕ_p is known as a hat function centred at mesh point p. A basis for $\mathcal{V}_h$ is given by $\{\phi_p : p \in \mathcal{N}\}$.

We introduce the mass lumping approximation using the quadrature rule:

$$\int_\Omega f \simeq \sum_{T \in \mathcal{T}_h} \frac{1}{3} A(T) \sum_{p \in \mathcal{N} \cap T} f(p). \tag{B.0.1}$$

The outer sum is over the triangles T in the triangulation, $A(T)$ denotes the area of the triangle T and the inner sum is over the three mesh points belonging to T. This rule is exact for $f \in \mathcal{V}_h$. It is well known that using this quadrature rule for the zero order terms in the piecewise linear finite element method for elliptic problems leads to no loss of order of accuracy in the energy norm as $h \to 0$ [20, Theorem 4.1.6].

This quadrature rule induces the discrete bilinear form:

$$\begin{aligned}\langle f,g\rangle &= \sum_{T\in\mathcal{T}_h}\frac{1}{3}A(T)\sum_{p\in\mathcal{N},p\in T}(fg)(p)\\ &= \sum_{p\in\mathcal{N}}w_p(fg)(p),\end{aligned}$$

where w_p equals a third of the sum of the areas of the triangles meeting at mesh point p. We can immediately see the benefit of this quadrature rule if we apply it to the term

$$(f,\phi_p) := \int_\Omega f\phi_p,$$

where ϕ_p is the piecewise linear basis function at mesh point $p \in \mathcal{N}$. (f,ϕ_p) is approximated by the discrete inner product:

$$\langle f,\phi_p\rangle = \sum_{q\in\mathcal{N}}w_q(f\phi_p)(q) = \sum_{q\in\mathcal{N}}w_q f(q)\delta_{pq} = w_p f(p).$$

So the approximation of (f,ϕ_p) only involves the value of f at mesh point p. This is particularly useful in the case of semilinear problems where f depends on the unknown solution of the PDE. $\langle f,\phi_p\rangle$ is said to be the **mass lumped** approximation to (f,ϕ_p).

The corresponding one dimensional nodal discrete inner product on the mesh $0 = x_0 < x_1 < \ldots < x_{n+1} = 1$ is given by

$$\langle f,g\rangle = \sum_{p=1}^{n+1}h_p\left\{\frac{f(x_p)g(x_p)+f(x_{p-1})g(x_{p-1})}{2}\right\}$$

where $h_p := x_p - x_{p-1}$, $p = 1,2,\ldots,n+1$.

Appendix C

Miscellaneous Finite Element Theory

In this appendix we prove a number of lemmas used in the proof of the *a posteriori* error estimates in Chapter 5.

First we make some limited assumptions on our problem, these fit in with the assumptions made in Chapter 5. Let $\Omega \subset \mathbb{R}^2$ be the polygonal domain we are working in, let $\partial\Omega_D$ be a non-empty subset of the boundary of Ω and assume $g \in \mathrm{H}^{\frac{3}{2}}(\partial\Omega_D)$ is value of the Dirichlet boundary condition defined on $\partial\Omega_D$. Define $\mathcal{V} = \mathrm{H}^1(\Omega)$ and $\mathcal{V}_g := \{v \in \mathrm{H}^1(\Omega) : v = g \text{ on } \partial\Omega_D\}$. We assume there exists a function $f(u(\boldsymbol{x}), \boldsymbol{x})$ such that $f : \mathbb{R} \times \Omega \to \mathbb{R}$ has the property that for all $\boldsymbol{x} \in \Omega$, $f(\cdot, \boldsymbol{x}) \in \mathcal{C}^2(\mathbb{R})$ and if $u \in \mathcal{C}(\Omega)$ then the function $\boldsymbol{x} \to f(u(\boldsymbol{x}), \boldsymbol{x})$ is in $\mathrm{L}_\infty(\Omega)$. For convenience denote $f(u(\boldsymbol{x}), \boldsymbol{x})$ by $f(u)$ for all $\boldsymbol{x} \in \Omega$, $f'(u)$ and $f''(u)$ will be taken to mean the derivatives with respect to u.

Assume there exists a $u_0 \in \mathcal{V}_g \cap \mathrm{L}_\infty(\Omega)$ such that

$$\mathrm{F}(u_0) = 0 \ \text{ in } \ (\mathcal{V}_0)', \tag{C.0.1}$$

where $\mathrm{F} : \mathcal{V} \to (\mathcal{V}_0)'$ is defined by

$$(\mathrm{F}(u), v) := (\nabla u, \nabla v) + (f(u), v), \quad u \in \mathcal{V}, \ v \in \mathcal{V}_0. \tag{C.0.2}$$

Further assume that F is continuously differentiable and has a bounded invertible

Fréchet Derivative, $\mathrm{F}' : \mathcal{V} \to L\left(\mathcal{V}_0, (\mathcal{V}_0)'\right)$, given by

$$(\mathrm{F}'(u)v, w) = (\nabla v, \nabla w) + (f'(u)v, w), \quad u \in \mathcal{V},\ v, w \in \mathcal{V}_0. \tag{C.0.3}$$

We assume that there exists constants Λ^0, γ_0, such that

$$\| \left(\mathrm{F}'(u_0)\right)^{-1} \|_{L((\mathcal{V}_0)', \mathcal{V}_0)} \leq \Lambda^0, \quad \|\mathrm{F}'(u_0)\|_{L(\mathcal{V}_0, (\mathcal{V}_0)')} \leq \gamma_0. \tag{C.0.4}$$

Finally assume for all $b \in \mathrm{L}_2$ there exists a unique $w \in \mathrm{H}^{1+\alpha}$ solving

$$\mathrm{F}'(u_0)\ w = b \ \text{ in } (\mathcal{V}_0)'$$

and

$$\|w\|_{1+\alpha} \leq \Lambda^0 \|b\|_0. \tag{C.0.5}$$

Let X be a given space with associated norm $\|\cdot\|$. Define the open X ball centred at $u \in X$ with radius r to be:

$$\mathcal{B}(u, r)_X \ := \ \{v \in X : \|u - v\| < r\}.$$

C.1 Bounding Lemmas

Lemma C.1.1 *For all $t \in [0, 1]$, there exists a constant $\tilde{C}$ such that for all r, r' sufficiently small:*

$$\|F'(u_1 + t(u_2 - u_1)) - F'(u_1)\|_{L(\mathcal{V}_0, (\mathcal{V}_0)')} \leq \tilde{C} t \|u_1 - u_2\|_0$$

when $u_1, u_2 \in \mathcal{B}(u_0, r)_{\mathrm{H}^1} \cap \mathcal{B}(u_0, r')_{\mathrm{L}_\infty}$.

Proof

In this proof we let sup denote the supremum over all $v, w \in \mathcal{V}$ with $\|v\|_1 = \|w\|_1 = 1$.

Thus

$$\begin{aligned} &\|\mathrm{F}'(u_1 + t(u_2 - u_1)) - \mathrm{F}'(u_1)\|_{L(\mathcal{V}_0, (\mathcal{V}_0)')} \\ &\qquad := \ \sup | \left(\left[\mathrm{F}'(u_1 + t(u_2 - u_1)) - \mathrm{F}'(u_1) \right] v, w \right) | \\ &\qquad = \ \sup |(\nabla v, \nabla w) + (f'(u_1 + t(u_2 - u_1))v, w) \end{aligned}$$

$$
\begin{aligned}
& -(\nabla v, \nabla w) - (f^{'}(u_1)v, w)| \\
= \quad & \sup|([f^{'}(u_1 + t(u_2 - u_1)) - f^{'}(u_1)]v, w)|.
\end{aligned}
$$

Using the generalised Hölders inequality and the Sobolev Imbedding Theorems [1] one obtains

$$
\begin{aligned}
& \|\mathrm{F}^{'}(u_1 + t(u_2 - u_1)) - \mathrm{F}^{'}(u_1)\|_{L(\mathcal{V}_0,(\mathcal{V}_0)')} \\
& \qquad \leq \sup \|f^{'}(u_1 + t(u_2 - u_1)) - f^{'}(u_1)\|_0 \|v\|_{\mathrm{L}_4} \|w\|_{\mathrm{L}_4} \\
& \qquad \leq \sup \|f^{'}(u_1 + t(u_2 - u_1)) - f^{'}(u_1)\|_0 \|v\|_1 \|w\|_1 \\
& \qquad = C\|f^{'}(u_1 + t(u_2 - u_1)) - f^{'}(u_1)\|_0. \qquad \text{(C.1.6)}
\end{aligned}
$$

By the mean value theorem there exists a $\theta_t(\boldsymbol{x}) \in \mathcal{B}(u_0, r)_{\mathrm{H}^1} \cap \mathcal{B}(u_0, r^{'})_{\mathrm{L}_\infty}$ between $u_1(\boldsymbol{x})$ and $u_1(\boldsymbol{x}) + t(u_2(\boldsymbol{x}) - u_1(\boldsymbol{x}))$, such that

$$
\begin{aligned}
\left[f^{'}(u_1 + t(u_2 - u_1)) - f^{'}(u_1)\right] &= f^{''}(\theta_t)([u_1 + t(u_2 - u_1)] - u_1) \\
&= f^{''}(\theta_t)t(u_2 - u_1). \qquad \text{(C.1.7)}
\end{aligned}
$$

From (C.1.6), (C.1.7) and the assumptions made on f, we deduce that there exists a $\tilde{C}$ such that

$$
\|\mathrm{F}^{'}(u_1 + t(u_2 - u_1)) - \mathrm{F}^{'}(u_1)\|_{L(\mathcal{V}_0,(\mathcal{V}_0)')} \leq \tilde{C}t\|u_1 - u_2\|_0.
$$

□

Using this lemma we can prove the following:

Lemma C.1.2 *Let $u_0 \in \mathcal{V}_g \cap L_\infty$ be the solution of $F(u) = 0$ in $(\mathcal{V}_0)'$. For r, r' sufficiently small:*

$$
(2\gamma_0)^{-1}\|F(u_1)\|_{(\mathcal{V}_0)'} \leq \|u_0 - u_1\|_1 \leq 2\Lambda^0\|F(u_1)\|_{(\mathcal{V}_0)'}, \qquad \text{(C.1.8)}
$$

for all $u_1 \in \mathcal{B}(u_0, r)_{\mathrm{H}^1} \cap \mathcal{B}(u_0, r^{'})_{\mathrm{L}_\infty}$.

Proof We begin by proving the right hand bound of (C.1.8).

We consider $u_1 \in \mathcal{B}(u_0, r)_{\mathrm{H}^1} \cap \mathcal{B}(u_0, r')_{\mathrm{L}_\infty}$, then

$$\begin{aligned} \mathrm{F}(u_1) - \mathrm{F}(u_0) &= \int_0^1 \frac{d}{dt} \{\mathrm{F}(u_0 + t(u_1 - u_0))\}\, dt \\ &= \int_0^1 \mathrm{F}'(u_0 + t(u_1 - u_0))(u_1 - u_0) dt. \end{aligned}$$

Thus

$$\mathrm{F}(u_1) - \mathrm{F}(u_0) - \mathrm{F}'(u_0)(u_1 - u_0) = \int_0^1 \left[\mathrm{F}'(u_0 + t(u_1 - u_0)) - \mathrm{F}'(u_0)\right](u_1 - u_0) dt. \tag{C.1.9}$$

Taking norms and using the result of Lemma C.1.1, we obtain:

$$\begin{aligned} &\|\mathrm{F}(u_1) - \mathrm{F}(u_0) - \mathrm{F}'(u_0)(u_1 - u_0)\|_{(\mathcal{V}_0)'} \\ &\quad \leq \int_0^1 \|\mathrm{F}'(u_0 + t(u_1 - u_0)) - \mathrm{F}'(u_0)\|_{L(\mathcal{V}_0, (\mathcal{V}_0)')} \|(u_1 - u_0)\|_1 dt \\ &\quad \leq \int_0^1 \tilde{C} t \|u_0 - u_1\|_0 \|u_0 - u_1\|_1 dt \\ &\quad \leq \int_0^1 \tilde{C} t \|u_0 - u_1\|_1^2 dt \\ &\quad \leq \frac{\tilde{C}}{2} \|u_0 - u_1\|_1^2, \end{aligned} \tag{C.1.10}$$

where $\tilde{C}$ is the constant from Lemma C.1.1.

Since we have assumed that u_0 solves (C.0.1), (C.1.10) tells us that

$$\|\mathrm{F}(u_1) - \mathrm{F}'(u_0)(u_1 - u_0)\|_{(\mathcal{V}_0)'} \leq \frac{\tilde{C}}{2} \|u_0 - u_1\|_1^2. \tag{C.1.11}$$

By assumption (C.0.5):

$$\| \left(\mathrm{F}'(u_0)\right)^{-1} g\|_1 \leq \Lambda^0 \|g\|_{(\mathcal{V}_0)'}, \quad g \in (\mathcal{V}_0)',$$

or equivalently

$$\|\mathrm{F}'(u_0) u\|_{(\mathcal{V}_0)'} \geq \left(\Lambda^0\right)^{-1} \|u\|_1, \quad u \in \mathcal{V}_0. \tag{C.1.12}$$

Therefore we may deduce, using (C.1.12), that

$$\|\mathrm{F}(u_1) - \mathrm{F}'(u_0)(u_1 - u_0)\|_{(\mathcal{V}_0)'}$$

$$
\begin{aligned}
&= \|\mathrm{F}'(u_0)\left[\left(\mathrm{F}'(u_0)\right)^{-1}\mathrm{F}(u_1) - (u_1 - u_0)\right]\|_{(\mathcal{V}_0)'} \\
&\geq (\Lambda^0)^{-1}\|\left(\mathrm{F}'(u_0)\right)^{-1}\mathrm{F}(u_1) - (u_1 - u_0)\|_1 \\
&\geq (\Lambda^0)^{-1}\left[\|u_1 - u_0\|_1 - \|\left(\mathrm{F}'(u_0)\right)^{-1}\mathrm{F}(u_1)\|_1\right]. \qquad \text{(C.1.13)}
\end{aligned}
$$

Thus, from (C.1.11) and (C.1.13):

$$
(\Lambda^0)^{-1}\left[\|u_0 - u_1\|_1 - \|\left(\mathrm{F}'(u_0)\right)^{-1}\mathrm{F}(u_1)\|_1\right] \leq \frac{\tilde{C}}{2}\|u_0 - u_1\|_1^2
$$

and rewriting this in a more useful form, we find that

$$
\|u_0 - u_1\|_1 \leq \|\left(\mathrm{F}'(u_0)\right)^{-1}\mathrm{F}(u_1)\|_1 + \frac{\tilde{C}\Lambda^0}{2}\|u_0 - u_1\|_1^2. \qquad \text{(C.1.14)}
$$

Since $u_1 \in \mathcal{B}(u_0, r)_{\mathrm{H}^1} \cap \mathcal{B}(u_0, r')_{\mathrm{L}_\infty}$, choosing $r \leq (\tilde{C}\Lambda^0)^{-1}$:

$$
\|u_0 - u_1\|_1 \leq (\tilde{C}\Lambda^0)^{-1}.
$$

We conclude from (C.1.14) that:

$$
\|u_0 - u_1\|_1 \leq \|\left(\mathrm{F}'(u_0)\right)^{-1}\mathrm{F}(u_1)\|_1 + \frac{1}{2}\|u_0 - u_1\|_1,
$$

which yields the upper bound of (C.1.8):

$$
\begin{aligned}
\|u_1 - u_0\|_1 &\leq 2\,\|\left(\mathrm{F}'(u_0)\right)^{-1}\mathrm{F}(u_1)\|_1 \\
&\leq 2\Lambda^0\|\mathrm{F}(u_1)\|_{(\mathcal{V}_0)'}
\end{aligned}
$$

as required.

We now prove the lower bound of equation (C.1.8):

Since we have assumed that u_0 solves (C.0.1) we may rearrange (C.1.9) and use Lemma C.1.1 to conclude that:

$$
\begin{aligned}
\|\mathrm{F}(u_1)\|_{(\mathcal{V}_0)'} &\leq \|\mathrm{F}'(u_0)\|_{L(\mathcal{V}_0,(\mathcal{V}_0)')}\|u_0 - u_1\|_1 \\
&\quad + \int_0^1 \|\mathrm{F}'(u_0 + t(u_1 - u_0)) - \mathrm{F}'(u_0)\|_{L(\mathcal{V}_0,(\mathcal{V}_0)')}\|u_0 - u_1\|_1 dt \\
&\leq \|\mathrm{F}'(u_0)\|_{L(\mathcal{V}_0,(\mathcal{V}_0)')}\|u_0 - u_1\|_1 + \int_0^1 \tilde{C}t\|u_0 - u_1\|_1^2 dt
\end{aligned}
$$

$$\leq \|\mathrm{F}'(u_0)\|_{L(\mathcal{V}_0,(\mathcal{V}_0)')}\|u_0 - u_1\|_1 + \frac{\tilde{C}}{2}\|u_0 - u_1\|_1^2. \tag{C.1.15}$$

We have assumed $u_1 \in \mathcal{B}(u_0, r)_{\mathrm{H}^1} \cap \mathcal{B}(u_0, r')_{\mathrm{L}_\infty}$, therefore taking $r \leq \left(2(\tilde{C})^{-1}\gamma_0\right)$:

$$\|u_0 - u_1\|_1 \leq \left(2(\tilde{C})^{-1}\gamma_0\right),$$

which, taken together with (C.1.15) and (C.0.4), implies the lower bound of (C.1.8):

$$\begin{aligned} \|\mathrm{F}(u_1)\|_{(\mathcal{V}_0)'} &\leq \gamma_0\|u_0 - u_1\|_1 + \gamma_0\|u_0 - u_1\|_1 \\ &= 2\gamma_0\|u_0 - u_1\|_1. \end{aligned}$$

□

Bibliography

[1] R.A. ADAMS. *Sobolev Spaces.* Academic Press, 1975.

[2] M. AINSWORTH. The performance of Bank-Weiser's error estimator for quadrilateral finite elements. Technical report, Mathematics Department, Leicester University, 1993.

[3] E. ALLGOWER, K. BÖHMER, and S. McCORMICK. Discrete defect corrections: Basic ideas. *ZAMM*, 62:371–377, 1982.

[4] E.L. ALLGOWER and K. BÖHMER. Application of the mesh independent principle to mesh refinement strategies. *SIAM J. Numererical Analysis*, 24:1335–1351, 1987.

[5] A. AMBROSETTI and G. PRODI. *A Primer of Nonlinear Analysis.* Cambridge, 1993.

[6] O. AXELSSON. On mesh independence and Newton-type methods. *Applications of Mathematics*, 38:249–265, 1993.

[7] I. BABUŠKA and W.C. RHEINBOLDT. Error estimates for adaprive finite element computations. *SIAM J. Numerical Analysis*, 15:736–754, 1978.

[8] R.E. BANK, T.F. CHAN, W.M. COUGHRAN, and R.K. SMITH. The alternate-block-factorization procedure for systems of partial differential equations. *BIT*, 29:938–954, 1989.

[9] R.E. BANK, W.M. COUGHAN, M.A. DRISCOLL, R.K. SMITH, and W. FICHTNER. Iterative methods in semiconductor device simulation. *Computer Physics Communications*, 53:201–212, 1989.

[10] R.E. BANK, A.H. SHERMAN, and A. WEISER. Refinement algorithm and data structures for regular local mesh refinement. In *Scientific Computing (Applications of Mathematics and Computing to the Physical Sciences) (ed. R.S. Stepleman)*, pages 3–17. North Holland, 1983.

[11] R.E. BANK and A. WEISER. Some a posteriori error estimators for elliptic partial differential equations. *Mathematics of Computation*, 44:283–301, 1985.

[12] E. BÄNSCH and K.G. SIEBERT. A posteriori error estimation for nonlinear problems by duality techiques. Technical report, Universität Freiburg, 1995.

[13] C. BERNARDDI and V. GIRAULT. A local regularization operator for triangular and quadrilateral finite elements. Technical Report 95036, Laboratoire d'Analyse Numerique, Université Pierre et Marie Curie, Paris, 1996.

[14] F. BORNEMANN, B. ERDMANN, and R. KORNHUBER. A posteriori error estimates for elliptic problems in two and three space dimensions. *SIAM J. Numerical Analysis*, 33:1188–1204, 1996.

[15] S.C. BRENNER and L.R. SCOTT. *The Mathematical Theory of Finite Element Methods.* Springer-Verlag, 1994.

[16] F. BREZZI, L.D. MARINI, and P. PIETRA. Numerical simulation of semiconductor devices. *Computer Methods in Applied Mechanics and Engineering*, 75:493–514, 1989.

[17] F. BREZZI, L.D. MARINI, and P. PIETRA. Two-dimensional exponential fitting and applications to drift-diffusion models. *SIAM J. Numerical Analysis*, 26:1342–1355, 1989.

[18] J.F. BÜRGLER, W.M. COUGHRAN, and W. FICHTNER. An adaptive grid refinement strategy for the drift-diffusion equations. *IEEE Transactions on Computer-Aided Design*, 1991.

[19] T.F. CHAN and T.P. MATHEW. Domain decomposition algorithms. *Acta Numerica*, pages 61–143, 1994.

[20] P.G. CIARLET. *The Finite Element Method for Elliptic Problems.* North-Holland, 1978.

[21] P. CLÉMENT. Approximation by finite element functions using local regularization. *Revue Française d'Automatique, Informatique et Recherche Opérationnelle*, 2:77–84, 1975.

[22] R.K. COOMER. *Parallel Iterative Methods in Semiconductor Device Modelling.* PhD thesis, School of Mathematical Sciences, The University of Bath, 1994.

[23] R.K. COOMER and I.G. GRAHAM. Massively parallel methods for semiconductor device modelling. *Computing*, 56:1–28, 1996.

[24] W.M. COUGHRAN, M.R. PINTO, and R.K. SMITH. Continuation methods in semiconductor device simulation. *J. Computational and Applied Mathematics*, 26:47–65, 1989.

[25] M. CROUZEIX and J. RAPPAZ. *On Numerical Approximation in Bifurcation Theory.* Masson/Spring-Verlag, 1989.

[26] P.M. DE ZEEUW. Nonlinear multigrid applied to a one dimensional stationary semiconductor model. *SIAM J. Scientific and Statistical Computing*, 13:512–530, 1992.

[27] K. DELJOUIE-RAKHSHANDEH and E.M. DEELEY. Error indication and adaptive refinement in semiconductor device simulation. *Int. J. Electronics*, 1988.

[28] J.E. DENNIS and R.B. SCHNABEL. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, 1983.

[29] M. DRYJA and W. HACKBUSCH. On the nonlinear domain decomposition method. Technical report, Christian-Albrechts-Universität zu Kiel, D-24098 Kiel, Germany, 1995.

[30] C.M. ELLIOTT and A.R. GARDINER. One dimensional phase field computations. In D. Griffiths and G. Watson, editor, *Proceedings of the Dundee Numerical Analysis Conference*, 1993.

[31] K. ERIKSSON, D. ESTEP, P. HANSBO, and C. JOHNSON. Introduction to adaptive methods for differential equations. *Acta Numerica*, pages 105 – 158, 1995.

[32] K. ERIKSSON and C. JOHNSON. An adaptive finite element method for linear elliptic problems. *Mathematics of Computation*, 50:361–383, 1988.

[33] Q. FAN, P.A. FORSYTH, J.R.F. McMACKEN, and W. TANG. Performance issues for iterative solvers in device simulation. *SIAM J. Scientific Computing*, 17:100–117, 1996.

[34] L. GIRAUD and R.S. TUMINARO. Domain decomposition algorithms for PDE problems with large scale variations. In J. Xu and D.E. Keyes, editors, *Proc. Seventh Int. Conf. on Domain Decomposition Meths.*, number 180 in Contemporary Mathematics, pages 205–210, Providence, 1994. AMS.

[35] P. GRISVARD. *Singularities in Boundary Value Problems.* Masson/Springer-Verlag, 1992.

[36] H.K. GUMMEL. A self-consistent iterative scheme for one-dimensional steady state transistor calculations. *IEEE Transactions on Electron Devices*, 11:455–465, 1964.

[37] W. HACKBUSCH. *Iterative Solutions of Large Sparse Linear Systems of Equations.* Springer-Verlag, 1994.

[38] P. HANSBO and M. LEVENSTAM. *FEMLAB Users Manual.* Chalmers University of Technology, Department of Mathematics.

[39] G. HEISER, C. POMMERELL, J. WEIS, and W. FICHTNER. Three-dimensional numerical semiconductor device simulation: algorithms, architectures, results. *IEEE Transactions on Computer-Aided Design*, 10:1218–1230, 1991.

[40] V. HUTSON and J.S. PYM. *Applications of Functional Analysis and Operator Theory.* Academic Press, 1980.

[41] J.W. JEROME. Consistency of semiconductor modeling: an existence/stability analysis for the stationary Van Roosbroeck system. *SIAM J. Appled Mathematics*, 45:565–590, 1985.

[42] C. JOHNSON. *Numerical Solution of Partial Differential Equations by the Finite Element Method.* Cambridge University Press, 1992.

[43] T. KERKHOVEN. A proof of the convergence of Gummel's algorithm for realistic device geometries. *SIAM J. Numerical Analysis*, 23:1121–1137, 1986.

[44] T. KERKHOVEN. A spectral analysis of the decoupling algorithm for semiconductor simulation. *SIAM J. Numerical Analysis*, 25:1299–1312, 1988.

[45] T. KERKHOVEN and J.W. JEROME. L_∞ stability of finite element approximations to elliptic gradient equations. *Numerische Mathematik*, 57:561–575, 1990.

[46] R. KORNHUBER and R. ROITZSCH. On adaptive grid refinement in the presence of internal or boundary layers. *IMPACT of Computing in Science and Engineering*, 2:40–72, 1990.

[47] R. KORNHUBER and R. ROITZSCH. Self-adaptive finite element simulation of bipolar, strongly reverse-biased pn-junctions. *Communications in Numerical Methods in Engineering*, 9:243–250, 1993.

[48] E. LEITNER and S. SELBERHERR. Three-dimensional grid adaption using a mixed-element decomposition method. In H. RYSSEL and P. PICHLER, editors, *Simulation of Semiconductor Devices and Processes*, volume 6, pages 464–467, 1995.

[49] P.A. MARKOWICH. A singular perturbation analysis of the fundamental semiconductor device equations. *SIAM J. Applied Mathematics*, 44:896–928, 1984.

[50] P.A. MARKOWICH. *The Stationary Semiconductor Device Equations.* Springer-Verlag, 1986.

[51] P.A. MARKOWICH, C.A. RINGHOFER, and C. SCHMEISER. *Semiconductor Equations.* Springer-Verlag, 1990.

[52] J.C. MEZA and R.S. TUMINARO. A multigrid preconditioner for the semiconductor equations. *SIAM J. Scientific Computing*, 17:118–132, 1996.

[53] W.F. MITCHELL. A comparison of adaptive refinement techniques for elliptic problems. *ACM Transactions on Mathematical Software*, pages 326–347, 1989.

[54] J.M. ORTEGA and W.C. RHEINBOLDT. *Iterative Solutions of Nonlinear Equations in Several Variables.* Academic Press, 1970.

[55] T.F. PENA, E.L. ZAPATA, and D.J. EVANS. Finite element simulation of semiconductor devices on multiprocessor computers. *Parallel Computing*, 20:1129–1159, 1994.

[56] C.P. PLEASE. An analysis of semiconductor pn junctions. *IMA Journal of Applied Mathematics*, 28:301–318, 1982.

[57] J. POUSIN and J. RAPPAZ. Consistency, stability, a priori and a posteriori errors for Petrov-Galerkin methods applied to nonlinear problems. *Numerische Mathematik*, 69:213–231, 1994.

[58] W. RUDIN. *Principles of Mathematical Analysis* . McGraw-Hill International Editions, 1989.

[59] M. SARANITI, A. REIN, G. ZANDLER, P. VOGL, and P. LUGLI. An efficient multigrid poisson solver for device simulations. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 15:141–150, 1996.

[60] A.H. SCHATZ, V. THOMÉE, and W.L. WENDLAND. *Mathematical Theory of Finite and Boundary Element Methods.* Birkhäuser, 1990.

[61] L.R. SCOTT and S. ZHANG. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Mathematics of Computation*, 54:483–493, 1990.

[62] S. SELBERHERR. *Analysis and Simulation of Semiconductor Devices.* Springer-Verlag, 1984.

[63] B. SMITH, W. GROPP, and L.C. McINNES. *PETSc 2.0 Users Manual.* Argonne National Laboratory, Mathematics and Computer Science Division.

[64] J. SMOLLER. *Shock Waves and Reaction Diffusion Equations.* Springer-Verlag, 1983.

[65] J.J. SPARKES. *Semiconductor Devices.* Chapman & Hall, 1994.

[66] G. STRANG. *Linear Algebra and its Applications.* Harcourt Brace Jovanovich, 1988.

[67] S.M. SZE. *Physics of Semiconductor Devices.* Wiley, 1981.

[68] W.V. VAN ROOSBROECK. Theory of flow of electrons and holes in gemanium and other semiconductors. *Bell Syst. Tech. J.* , 29:560–607, 1950.

[69] R.S. VARGA. *Matrix Iterative Analysis.* Prentice-Hall International, 1962.

[70] R. VERFÜRTH. A posteriori error estimates for nonlinear problems. Finite element discretization of elliptic equations. *Mathematics of Computation*, 62:445–475, 1994.

[71] R. VERFÜRTH. A posteriori error estimates for nonlinear problems. L_r-estimates for finite element discretizations of elliptic equations. Technical Report 199, Ruhr-Universität Bochum, 1996.

[72] J. XU. A novel two-grid method for semilinear elliptic equations. *SIAM J. Scientific Computing*, 15:231–237, 1994.

[73] J. XU. Two-grid discretization techniques for linear and nonlinear PDEs. *SIAM J. Numerical Analysis*, 33:1759–1777, 1996.

[74] D.M. YOUNG. *Iterative Solution of Large Linear Systems.* Academic Press, 1971.

[75] H. YSERENTANT. On the multi-level splitting of finite element spaces. *Numerische Mathematik*, 49:379–412, 1986.